Molecular Diversity and Functional Composition of Cellulose Degrading Communities in Anoxic Environments

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

James Nicholas Ian Houghton

September 2013

NATURAL ENVIRONMENT RESEARCH COUNCIL

UNIVERSITY OF LIVERPOOL

Abstract

The major fraction of microbial communities cannot be cultivated by artificial means in the laboratory. In order to access the full diversity of microbial life in the open environment it is necessary to employ culture independent methods. Molecular biology and now metagenomics have enabled the phylogenetic and functional investigation of microbial communities without isolation and cultivation of organisms and has led to a new appreciation of the breadth of diversity of microbes on Earth and to the discovery and characterisation of new enzymes. Here, molecular biological techniques have been applied to the study of microbial communities specifically in anaerobic environments and with an emphasis on those involved in the primary degradation of plant cellulosic biomass. Quantitative PCR was used to assess the presence of cellulolytic bacteria both in landfill leachate and specifically in association with cotton cellulose “baits” maintained in leachate microcosms. Lineages of clostridia previously associated with cellulose degrading strains were detected in all five of the landfill leachate samples, and *Fibrobacter spp.* were detected at low abundance (2.3% of total bacteria) in one sample. Clostridia Group III and *Fibrobacter spp.* were enriched on the surface of a bait (17% and 29% of total bacteria, respectively) that was rapidly degraded by the colonising community and were present in low abundance (< 1%) and absent, respectively, on another colonised by a community which did not exhibit any degradation of the cellulose. The observed correlation between high levels of cellulose degradation and presence *Fibrobacter spp*. demonstrates a cellulolytic role outside of the gut environment for these organisms the first time.

A metatranscriptome was prepared from a set of cotton cellulose baits maintained in a lake sediment for 2-8 weeks, and Illumina sequencing was used to generated ca. 7 million paired-end reads. Just under one million putative protein coding sequences were identified and of these, MEGAN analysis determined that 40% had no blast hit to the NCBI NR database suggesting that a large number of unknown sequences were present. Analysis of this metatranscriptome and a metagenome produced from the same site revealed that bacteria accounted for 75% of the protein coding sequences and 97% of the metagenome. Genes with matches to cellulolytic lineages of clostridia were found to be present and *Fibrobacter* sequences were also detected in both of these datasets further demonstrating their presence in the wider environment as probable cellulose degraders

ORF prediction and HMM searching were used to search for expressed cellulases in the metatranscriptome and identified 503 sequences with high similarity to glycoside hydrolase protein families, representing carbohydrate active enzymes with possible cellulolytic activity. Of these 112 were also found to have representatives in the metagenome with 100% sequence similarity. All of these sequences had a low level of identity to entries in the NCBI NR database indicating the discovery of previously unknown genes.

A fosmid library was produced from the same DNA used to generate the metatranscriptome and it is possible that full-length copies of the expressed genes identified *in silico* will have been captured. This fosmid library can be interrogated accordingly using probe and PCR primer sequences designed using the curated metatranscriptome dataset. In this way, potentially novel cellulases can be discovered for biochemical characterisation, genetic manipulation and biotechnological exploitation.

Acknowledgements

Thanks to my supervisors Prof. Alan McCarthy and Dr Heather Allison for advice, support and the occaisional pep talk throughout my PhD. I am grateful to them and NERC for the opportunity to undertake my PhD studentship. Thanks also to Dr Mal Horsburgh who is partly redsponsible for persuading me that a PhD was a good idea in the first place and has throughout been a source of advice, albeit of a somewhat sarcastic nature. Paul Loughnane deserves credit for having the patience of a saint and has put up all manner of questions and queries during my time in the lab.

Dr James McDonald not only wrote the grant that funded much of my work but also kindly provided DNA and landfill leachate samples on which many of my first year experiments were based. I am also greatful for the advice and instruction on the finer points of quantitative PCR, which I now found myself passing on to others in turn - you're time wasn't wasted.

Dr Dave Rooks has been a supportive postdoc during the project whose knowledge and guidance have kept me on the straight and narrow, whose temper tantrums have kept me greatly amused and has left we with some great memories, like being towed out of a field by an enormous dumper truck that time.

I am grateful for the camaraderie of all those from lab H and around the institute generally who have experienced the trials and tribulations of scientific reasearch with me and all of the highs and lows that come with it. To everyone who I have shared accomodation with over the past four years, or has had a drink with me in the AJ, thanks for the company and quite probably for putting up with me moaning when I was either making a hash of things in the lab or things were making a hash of themselves.

Finally, thanks to all of my family, in particular my parents who have helped me get this far and a provided a great deal of financial help along the way and the Lancashire Houghton clan for all of their hospitality over the past few years. Huge thanks also to Jen who made the months of writing up much less irritating and mind-numbing, and was (mostly) very patient with me.

Table of Contents

Abbreviations

BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
bp	Base pair
CAZy	Carbohydrate active enzymes
cDNA	Complementary DNA
CMC	Carboxy methyl cellulose
Ct	Threshold Cycle
CTAB	Cetyltrimethylammonium bromide
DEPC	Diethylpyrocarbonate
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
dscDNA	Double-stranded complementary DNA
PFGE	Pulse-Field Gel Electrophoresis
EDTA	Ethylenediaminetetraacetic
HMM	Hidden Markov Model
HMW	High molecular weight
LB	Luria Bertani medium
LSU	Large Subunit
mRNA	Messenger RNA
NCBI	National Centre for Biotechnology Information
NR	Non-redundant (i.e. Non-reduant protein database)
ORF	Open Reading Frame
OUT	Operational Taxanomic Unit
PEG	Polyethylene glycol
PCR	Polymerase Chain Reaction
PolyA	Poly Adenine
qPCR	Quantitative PCR
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
sddH_20	Sterile double distilled H_20
SDS	Sodium Dodecyl Sulphate
ssDNA	Single stranded DNA

SSU	Small Subunit
SET	Sucrose-EDTA-Tris
STE	Saline-Tris-EDTA
TAE	Tris-acetate-EDTA
TBE	Tris-borate-EDTA
TE	Tris-EDTA
TGGE	Temperature Gradient Gel Electrophoresis
T_m	Melting Temperature
Tris	Tris(hydroxymethyl)aminomethane
v/v	Volume/Volume
w/v	Weight/Volume

IUPAC Degenerate Base Symbols

A	**A**denosine
C	**C**ytidine
G	**G**uanosine
T	**T**hymidine
U	**U**racil
W	**W**eak (A or T)
S	**S**trong (G or C)
M	A**m**ino (A or C)
K	**K**eto (G or T)
R	Pu**r**ine (A or G)
Y	P**y**rimidine (C or T)
B	Not A
D	Not C
H	Not G
V	Not T
N	Any base (or unspecified base)

Chapter One: Introduction

1.1 Background: Global microbial abundance.

The global abundance of microbial cells is huge. Microorganisms are the dominant form of life, certainly numerically and in terms of their sheer diversity. It has become apparent that individual species of bacteria in complex communities such as those found in soil are numbered in thousands (Roesch *et al*. 2007; Urich *et al*., 2008). Furthermore, microbial life permeates the most extreme environments on the planet including desiccated deserts, hot springs, deep sea thermal vents and the frozen polar regions. Microbial life drives essential biogeochemical cycles and effectively permits the existence of other life forms.

The range of 4 – 6 $x10^{30}$ was put forward as being the estimated number of prokaryotic cells globally by Whitman *et al.* (1998). It has been subsequently demonstrated by Kallmeyer *et al*. (2012) that this number may be inflated and deep seafloor sediment microbial abundance in particular is probably less than originally projected. Putting a number on the global quantity of prokaryotes is a difficult task and a considered estimate, which requires updating after scientific and technological advances of more than a decade, is probably not unreasonable. The central point raised by Whitman *et al.* is left firmly intact by the investigations of Kallmeyer *et al*. and is simply that there are an enormous number of microorganisms in the world, and they are an extremely important part of the biosphere.

There is therefore an interest in studying microbial communities in the environment from the point of view of understanding their contribution to biogeochemical cycles, their ability to break down and release energy and nutrients otherwise imprisoned in the dead organic matter of higher organisms. One other reason to investigate environmental microorganisms is that they represent a huge potential resource of genetic material encoding specialised enzymes that can be harnessed for industrial applications. There is the potential to unlock novel biology, and perhaps even provide solutions to the problems of renewable energy generation, bioremediation or the search for new antimicrobial agents (Ferrer *et al*., 2005; Li *et al*. 2009). Investigating these lines of enquiry is challenging, if true novelty is to be discovered.

1.2 Molecular microbial ecology

The field of study that can be called molecular microbial ecology consists of methods developed to overcome the inherent limitations of the conventional

approach to microbiology, as well as systems for analysing and interpreting the data generated by these methods. Classical microbiology is culture-based; the isolation of organisms for production of pure cultures, which in turn enables the examination of their morphology, physiology, biochemistry and genetics.

1.2.1 The culturing problem.

Isolation of microorganisms as pure colonies using traditional culturing methods is at best an incomplete method of surveying environmental microbial communities. Obtaining a discrete colony on an agar plate which can be described phenotypically and propagated for nucleic acid extraction and further molecular analysis is a robust way of discovering and characterising new species, but is simply insufficient for dealing with the microbiota of the biosphere since the majority of bacteria present cannot be cultivated in the laboratory. Culturing techniques are thought to capture at best 10% of the communities carried by animals and only 1% or less of the species found in the wider environment (Bomar *et al*., 2011). Culturing techniques can be tailored in an attempt to accommodate the growth requirements of microorganisms adapted to specific conditions, and cultivation rates of 10-20% of organisms present in environments such as freshwater lakes and marine sediments are reportedly attainable (Teeling & Glöckner, 2012). This still leaves a large proportion of the microbiota inaccessible however, and many species may be exceptionally hard to cultivate if, for example, they are strictly anaerobic or dependant on a symbiotic relationship with other organisms.

The reasons for this are that environmental communities of microorganisms are large heterogenous assemblages of microbes that are likely to be interdependent, have their own specific growth requirements and may have adapted to states of nutrient limitation. Attempting to promote the growth of such organisms in the laboratory successfully could require very precise duplication of the conditions of the microenvironments to which individual organisms have become adapted. Parameters include pH, specific availability of certain nutrients and specific oxygen concentrations. Consider that a tiny grain of soil could harbour aerobes on the surface, microaerophiles in internal crevices and perhaps strict anaerobes in internal cavities. Attempting to isolate one set of these organisms would automatically preclude the selection of the others. Using culturing techniques may well lead to the identification of novel and interesting species from environmental samples but will be heavily biased, delivering only a limited snapshot of a handful of amenable constituents and will be useless for understanding the community as a whole or obtaining any sort of complete survey.

Fortunately, the birth and development of molecular biology has furnished the scientific community with tools that enable microbial communities to be investigated in a more comprehensive fashion. The application of molecular techniques to environmental microbiology has evolved with time, and continues to do so. The range of techniques available is large and choosing a particular methodology is dependent on the goal of the experiment. Almost all have required the extraction of total community nucleic acid from environmental samples as the starting point, although single cell genomics is now beginning to be feasible (Rinke *et al*., 2013). Proteomics and its application to microbial ecology as metaproteomics is gaining momentum and has been used in some studies, but the majority of microbial ecology work focuses on genes and genomes.

1.2.2 Sampling for molecular microbial ecology

Extraction of nucleic acids from environmental samples can be performed in a number of different ways and will depend upon the nature of the sample and the intended application. Whether DNA or RNA or both are required, fragment size needed, the chemical content of the sample and its physical state will all influence the choice of extraction process. (Zhou *et al*., 1996; Rajendhran & Gunasekaran, 2008). There is a particular challenge for environmental microbiologists as contamination of nucleic acid preparations is an issue, especially when working with soils and sediments which have high levels of organic matter. Humic substances bind DNA and RNA, are difficult to remove from samples but must be depleted as far as possible as their presence may well hinder downstream enzymatic reactions such as PCR (Bonot *et al*., 2010; Mettel *et al*., 2010; Griffiths *et al*., 2000). There is also the issue of ensuring that nucleic acid preparations constitute as complete and even a representation of the community as possible. Cells buried in biofilms or deep within soil particles are likely to suffer less lysis in the extraction process than more accessible cells. Unless care is taken at this early step of sample processing, biases are likely to be introduced which may not become apparent until data analysis is performed later (He *et al*. 2009).

Generally, there are two main processes involved in extracting the nucleic acids from environmental samples. Chemical lysis involves the use of buffer solutions containing components which are able to break open the envelopes of cells and thereby free the DNA and RNA molecules within. The composition of buffers and concentrations of the components are variable and often tailored to particular protocols and sample types, but typical ingredients are detergents such as sodium dodecyl sulphate (SDS), lysozyme, the chelating agent EDTA, Tris and CTAB.

These substances in combination are able to release nucleic acids and offer a measure of protection against enzymatic damage from nucleases present in the sample.

Mechanical lysis employs the use of small glass, or occasionally metal, beads. Samples are placed into tubes containing beads which are shaken mechanically at extremely high speeds in a ribolyser or similar apparatus, resulting in a homogenised sample (Rajendhran & Gunasekaran, 2008). Mechanical and chemical lysis methods are frequently employed in combination to obtain a synergistic effect (Rosewarne *et al*., 2011; Yu *et al*., 2004). The choice of lysis buffers, and mechanical disruption will always depend heavily on the nature of the sample and the intended application.

After incubation and treatment, pre-processing the nucleic acids must be purified out from the rest of the material contained in the sample. A plethora of commercial kits have been developed, engineered specifically to remove the contaminants associated with soil, stool samples, blood and animal tissue among others. Phenol:chloroform extraction is a traditional and effective method for nucleic acid cleanup, but the toxicity of these chemicals is severe and column-based cleanup kits are increasingly used wherever possible.

Ultimately, it is unlikely that nucleic acid extraction from an environmental sample will result in a completely even sampling of the genomic content. It is to be expected that some species will be over or under represented in the final dataset and rare species may be missed (Hong *et al*. 2009, Amend *et al*. 2010). Utilising multiple extraction methods on the same sample and then pooling DNA obtained from those multiple methods is one way to improve the evenness of sampling. If the sample material is limited in quantity however, this may not be feasible.

A Summary of chemical, physical and enzymatic treatments used in nucleic acid extraction from microbial cultures and environmental samples is presented in Table 1.1.

Table 1.1 Treatments used for the extraction of pure nucleic acids from cells for use in molecular biology applications.

Treatment/buffer component	Type of treatment	Description
Lysozyme	Enzymatic	Attacks the cell walls of Gram positive bacteria
Proteinase K	Enzymatic	Frequently employed to reduce protein contamination in nucleic acid preparations
CTAB (cetyltrimethylammonium bromide)	Chemical	When used in lysis buffers for nucleic acid extraction, reduced contaminating humic substances and polysaccharide material; also acts as a surfactant.
PVPP (Polyvinylpolypyrrolidone)	Chemical	Used for reducing contaminating humic substances in extractions
Phenol/choloroform*	Chemical	Denatures proteins; removes them to the organic phase after centrifugation leaving nucleic acid in the aqueous phase to be removed by pipetting or decanting.
SDS	Chemical	Detergent, disrupts cell membranes
Bead beating	Physical	Disrupts cell membranes
Heating	Physical	Increases activity of enzymatic and physical components, or at higher temperatures causes cell lysis and protein inactivation.

*Often used in the form of a tris-buffered phenol/chloroform/isoamyl alcohol solution at a ratio 25:24:1.

1.2.3 Culture independent analysis of microbial communities.

Total community nucleic acid extracts yield either DNA representing the genomic material of the community (a metagenome) or RNA representing the ribosomal, transfer and messenger RNA molecules of the entire community. When working with RNA it may be necessary to employ some method of retrieving the fraction of interest. Ribosomal RNA can be size fractionated for a taxonomic analysis or mRNA can be isolated to enable analysis of the community gene expression in the form of a metatranscriptome. RNA samples generally require conversion to cDNA for analysis as PCR and sequencing methods can only be applied to DNA.

The polymerase chain reaction is an efficient and accurate method for assessing metagenomic DNA. Consultation of sequence databases allows for the design of highly specific primers which can be used to screen for organisms at the genus or even species level, or target specific genes. Amplification of target genes in a sample can provide simple presence/absence indications, but PCR can be used

for targeted cloning of specific genes and to generate quantitative data. PCR can be used to generate amplicons of, for example, the 16S and 18S rRNA genes of Bacteria, Eukarya and Archaea. These amplicons can form the basis for clone library production and high-throughput sequence analysis to study the phylogeny of a microbial community. Apart from PCR amplification, studies of microbial communities can also be carried out using oligonucleotide probing techniques.

1.2.3.1 Cloning

Two types of cloning have been utilised with respect to environmental microbiology; shotgun cloning and cloning of PCR amplicons. The generation of clone libraries of particular PCR amplification products has been extensively used as a method of investigating the composition of microbial communities. This method is prone to many biases and inaccuracies, so this particular aspect of cloning is beginning to be surpassed by high throughput sequencing for conducting surveys of microbial communities (Hong *et al*. 2009).

Production of clones has one absolute advantage in that it produces actual amplifiable and analysable inserts, as opposed to just sequencing output data. Cloning into large vectors such as fosmids, cosmids or Bacterial artificial chromosomes (BACS) remains a useful technique. Fosmids can accommodate inserts in the region of 35-50 kb and these large fragments of DNA can contain entire protein coding sequences or even whole intact bacterial operons. The resulting clones can be used in expression screening assays to detect any enzyme activity present that is able to catalyse a specific process. Utilising shotgun cloning of metagenomic DNA in this way can lead to the discovery of novel enzymes. The ability to acquire entire genes is used to obtain large quantities of their protein products for further work and characterisation (Palackal *et al*. 2007). Fosmids, Cosmids and BACS, due to their size, are also useful for studying the genomic context of genes identified through expression screening; sequencing these large cloning vectors could reveal related, accessory or regulatory genes related to the function of the active enzyme or enzymes discovered during the screening process.

Cloning methods do possess pitfalls. The metagenome of a diverse and populous microbial community is of course large and there is a randomness to the cloning process which means it is difficult, or impossible, to guarantee that the specific genes which are being sought will appear in the resulting library, particularly in the case of shotgun cloning. Cloning of specific amplicons is affected by bias introduced by PCR, and both types of cloning are subject to their own biases. These flaws are problematic and have led to cloning being supplanted as a sensible way of

studying a community but the technique is still important for isolating novel genes, and, therefore, proteins.

1.2.3.2 Molecular Fingerprinting Techniques

Fingerprinting techniques have been applied to community DNA extractions as a way of obtaining a pattern demonstrating community complexity and as a reference point from which shifts in community composition and structure can be monitored. In this way, DNA fingerprinting has aided in monitoring the response of microbial populations to changing conditions and enrichments (Deng *et al*. 2008). Several methods of fingerprinting have been developed, including Denaturing Gradient Gel Electrophoresis (DGGE), Temperature Gradient Gel Electrophoresis (TGGE) and restriction fragment length polymorphisms (RFLP).

DGGE and TGGE are based on the amplification of the 16S rRNA genes from community DNA samples by PCR, or the reverse transcription of community RNA into cDNA as a template for 16S rRNA gene amplification. Amplification is performed with primers incorporating a GC-clamp which prevents full denaturation of dsDNA molecules. Amplicons are run on a gel where there is a gradient of either temperature (TGGE) or a denaturing agent (DGGE) and are separated based on the partial dissociation of the amplicons, as the mobility of individual sequences will vary across the gradient. This results in a pattern of bands, the intensity of which provide semi-quantitative data indicating how different groups might respond to changing conditions over time (Shin *et al*. 2008) or which groups might be adapted to specific environments through the analysis of different samples or treatment conditions (e.g. Nicol *et al*. 2007). Gray *et al*. (2003) used TTGE analysis to track the response of microbial communities to sewage sludge and lime applications to soil plots over time and it was demonstrated that although differences between community structures emerged due to spatial and temporal variation to a great extent, treatment did have effects on the microbial community structures with certain groups, such as ammonia-oxidising bacteria, responding more than others.

RFLP has also been used in metagenomic studies (Deng *et al*. 2005) although this technique has been somewhat superseded by the widespread availability of low-cost DNA sequencing technology. A variant of the technique called Terminal-RFLP (T-RFLP) still finds regular usage today however. This technique employs the use of PCR amplification with fluorescently labelled primers, and digestion of the resulting amplicons with a restriction enzyme. The resulting digestion can be visualised through the detection of the fluorescent dye and this technique was recently used by Zumsteg *et al*. (2012) to investigate the microbial colonisation

and community development in a glacier forfield. As ground is gradually laid bare by retreating glaciers, microbes colonise the surface and communities develop over time. Sampling of soils in the ground in the path of the retreating glaciers over several distances from the front reveals a temporal pattern of microbial colonisation and community development, in this case with cyanobacteria as first colonisers of bare rock, and proteobacteria becoming numerous inhabitants of recently bared surfaces. As surfaces become older and more vegetated further away from the glacier front these groups were found to decline. An initial colonisation of *Euryarchaeota* was found to give way to *Crenarcheota* in the case of the Archaea. T-RFLP was also used to study perturbation in forest soil microbial communities in response to monomethyl-mercury contamination (Rieder & Frey, 2012), and the response of soil microbes to land utilisation (Ying *et al*. 2013). The continuing use of T-RFLP is perhaps a pragmatic one; high throughput sequencing is not inexpensive and if multiple comparison samples of microbial communities are needed for a study, as in the ones cited above where multiple soil samples were compared to one another, it represents a cost-effective method to which statistical analysis can be applied.

There are limitations to fingerprinting techniques; extrapolation from band patterns on a gel is only partially informative. Unlike cloning, which has irreplaceable functionality, as sequencing becomes cheaper and analysis pipelines more automated and user-friendly, fingerprinting methods are likely to become redundant. Amplicon sequencing has already become a viable and accessible alternative to these methods.

1.2.3.3 Quantitative PCR

Quantitative PCR (qPCR) has been used to investigate microbial communities and obtain abundance data, moving beyond a simple presence/absence PCR assay and providing information about copy numbers of particular genes, and the proportion of the community target taxa represent. Applied in the correct context, qPCR has the potential to produce illuminating data, providing insights into the structure of microbial communities and the relationship between bacteria and their environment. The caveat here is that qPCR does demand a rigorous experimental design in order for meaningful data to be generated. There is a requirement for carefully controlled experiments, technical replication and good lab technique.

Like conventional PCR methods, there is of course bias associated with the data produced from this technique. Smith *et al*. (2006) demonstrated that there is a

particular problem of very high levels of variability between experiments when using qPCR to determine absolute numbers of specific gene or transcripts. As a consequence, it is advisable to express qPCR results as relative abundance data, as a fraction of the community examined in a given assay, to avoid inaccuracies and to produce data that can be compared between individual assays with confidence.

1.2.3.4 Molecular Probing

Probing as a method of detecting specific phylogenetic groups requires the design of taxon-specific oligonucleotide sequences. Some level of *a priori* knowledge of the community or communities to be studied, or at least a hypothesis regarding the likely composition, is necessary for making use of this technique so that appropriate sequences can be designed. Design of molecular probes based on knowledge gleaned from clone library content has been used as an effective way of studying the abundance of taxonomic groups (Daly & Shirazi-Beechey, 2003). Quantitative information can also be achieved through qPCR and high-throughput methods but molecular probing does have one unique feature which renders it particularly interesting. Fluorescent in-site hybridization (FISH) combines probing and microscopy to visually pinpoint the location of the target organisms. FISH has been used to demonstrate the specific adherence to cellulose in a bioreactor by clostridia initially detected via clone libraries which demonstrated that these organisms were involved in the decomposition of this substance and not just detected incidentally (Burrell *et al*. 2003). FISH can be a very powerful, and informative, technique that reveals information not attainable with cloning, sequencing or amplification.

1.2.3.5 High-throughput sequencing platforms for metagenomics.

The technological advances in molecular biology, physics and engineering have been combined effectively to produce huge amounts of sequencing data, and the output from high throughput sequencing technology has continued to increase in terms of read length, quality and total amount of information produced per run. The ability to sequence genetic material on a technological level has in many respects outpaced resources and methods for handling and analysing the data *in silico*.

Analysis of sequencing data derived from a sample of a single organism in culture grown in laboratory media, and where there is a reference genome available, is relatively straight-forward given a sufficiently powerful computer to run the necessary software. Analysis of metagenomic and metatranscriptomic data remains an extremely computationally demanding task, becoming increasingly difficult if the community that is being sequenced is very heterogeneous (Prakash *et al*. 2012).

Simpler communities may be encountered in, for example, hot springs where there is a very strong selective pressure so that only relatively highly adapted organisms can survive. Seawater, soils and sediments by comparison tend to contain populations of cells comprised of a very large range of species (Thomas *et al*. 2012).

There are multiple sequencing technologies available that extend beyond the classic Sanger sequencing method (Sanger *et al*. 1977) applied almost universally until recently. 454 Pyrosequencing has been widely used in metagenomic projects in the recent past (Sogin *et al*. 2006, Frias-Lopez *et al*. 2008, Luo *et al*. 2012). The reason for this is that it produces longer read-lengths than other systems, up to a maximum read length of approximately for 1000 bp, and this can facilitate assembly or assignment of an identity to an individual read. Recently, though, Illumina sequencing has developed improvements in read-length (see Table 1.2 below), and this continues to increase. In addition Illumina sequencing chemistry produces a much lower error rate associated with A/T rich homopolymer sequences, and is more cost effective, producing more reads and therefore more overall data for the same cost as can be achieved using 454 pyrosequencing (Luo *et al*. 2012). As a result Illumina sequencing is becoming much more attractive for metagenomic studies.

Table 1.2 presents a summary of currently available sequencing technology employed for metagenomic purposes.

Table 1.2 Summary of current commercially available sequencing technologies, adapted and updated from Glen (2011). This list is not exhaustive; other technologies exist but are a niche market, not routinely used for molecular ecology or are still in a development stage.

Instrument	**Millions of reads per run (Approximate)**	**Bases per read (Average)[a]**	**Data yield per run (MB)**
3730xl (capillary)	0.000096	650	0.06
454 FLX+	1	650	650
454 GS Jr. Titanium	0.1	400	50
Illumina Miseq V.1	4	150 + 150[b]	1200
Illumina Miseq V.2	15	250 + 250[b]	7500
Illumina HiSeq 1000	1500	100 + 100[b]	300, 000
Illumina HiSeq 2000	3000	100 + 100[b]	600, 000
Illumina Genome Analyzer II	300	150 + 150[b]	95, 000
Ion Torrent – '314' chip	0.1	400	40
Ion Torrent – '318' chip	4	400	1500
Ion Torrent – Proton I	70	200	10, 000

[a]Average read lengths given – longer reads are possible from 454 and Ion Torrent technology

[b] The + symbol indicates that paired-end sequencing can be performed using these machines, where a single read is sequenced from both ends, sometimes overlapping in the middle, giving two reads per sequence. Software exists to merge paired reads and create a single longer sequence.

Metagenomic studies use either a shotgun or amplicon approach. Amplifying, with PCR, a variable region of the 16S rRNA gene for Bacteria and Archaea, 18S rRNA gene for the Eukaryota or the ITS region in cases where the 16S or 18S gene will not yield sufficient resolution and then performing sequencing on these amplicons can provide an in-depth and detailed overview of a community structure with information concerning the relative abundance of various species. Alternatively, DNA extracted directly from environmental samples can be used to create a shotgun library of sequences. A shotgun library contains information regarding gene function in addition to phylogenetic information and is an insight into the metabolic potential of a population based on homology of the metagenomic reads to protein database sequences. The caveat remains that the majority of putative protein coding sequences in a metagenome will be of unknown origin and function. It can sometimes be possible to perform assembly of metagenomic datasets and obtain larger contigs representing entire genes or operons or larger regions, but this is dependent on the depth and coverage of sequencing. Analysis of individual shotgun reads can still provide some phylogenetic overview and insight into the metabolic potential of an environmental microbial community.

1.2.3.6 Metatranscriptomics

Metatranscriptomics is the investigation of the combined community gene expression of a microbial community by sequencing a total community mRNA sample. Sequencing of RNA libraries on high-throughput platforms is frequently referred to as RNA-seq although this is something of a misnomer as these technologies can only sequence DNA, and the RNA sample must first be converted to a cDNA library. This, and other issues, render the examination of a microbial community particularly time-consuming, expensive and technically challenging.

The first major problem to consider when investigating community gene expression through metatranscriptomics is that the majority of the RNA in cell is in fact ribosomal. Actual mRNA sequences typically make up no more than between 1 and 5% of the total (Moran *et al*., 2013). Simply converting total RNA to cDNA and sequencing the resulting fragments will yield an output consisting of mostly rRNA. Working on eukaryotic systems has the advantage that the natural polyadenylation of mRNA transcripts enables easy selection of these sequences, but microbial mRNA transcripts have no polyA tails and cannot be easily isolated away from the bulk of the RNA sample. Some studies have been carried out in which total community RNA samples were used to generate libraries for high-throughput sequencing, and the output from the sequencing step divided for processing by two

separate pipelines, one for rRNA reads and the other for putative mRNA sequences (Urich *et al*. 2008, Radax *et al*. 2012). This approach works if the intended goal of the investigation is to achieve both an understanding of the taxonomic community structure and an insight into the most expressed genes, and most important metabolic functions in a habitat. If the focus is on community gene expression however, then library production from total RNA with a resulting dataset containing a large number of unwanted rRNA reads will not be ideal and will waste a great proportion of sequencing output. Some form of rRNA reduction during sample processing will need to be implemented.

Methods exist for removal of rRNA sequences from total RNA extractions via subtractive hybridisation and 5'-monophosphate-dependant exonuclease activity, and both are effective at reducing total rRNA sequence quantity in an RNA sample (Mader *et al.*, 2011). Limitations arise due to the subtractive hybridisation probes not being efficient; the 5'-monophosphate-dependant exonuclease approach appears to target mRNA transcripts, a fraction of which are likely to be 5'-monophosphorylated (Mettel *et al.*, 2010).

He *et al*. (2010) performed an in-depth investigation into the performance of rRNA removal procedures and assessed not just the level of rRNA removal achieved but also monitored the effect such treatments had on the abundance of mRNA transcripts in the samples. This is an important consideration since although it is important to remove excess rRNA, the influence of sample processing on the mRNA reads introduces bias and impacts the conclusions eventually drawn from the data. Hybridisation was found to be the best removal option, leading to loss of a greater proportion of rRNA from processed samples, with no effect on mRNA transcript abundance. Exonuclease treatment did have an effect on transcript abundance and it was hypothesised that this was likely to be due to the use of 5'-monophosphate dependant exonucleases, relying on the enzymes discriminating between 5′-triphosphates on mature mRNA transcripts and 5'-monophosphates on mature rRNA molecules, so partially degraded mRNA transcripts lacking a 5′-triphosphate will be susceptible to the activity. The consequence of this, as pointed out by the authors, is that although this exonuclease treatment will have an effect on detectable transcript numbers, this method will result in detection of full-length mRNA molecules providing an indication of stable gene expression in a cell. Hybridisation treatment is however more likely to preserve the transcripts of genes that are being turned over more rapidly.

One caveat to bear in mind concerning the work of He *et al*. (2010) is that their study focussed only on artificial communities made up of 5 species, all of which

were culturable and defined isolates. Real communities with a more complex structure are likely to be harbouring novel or less well studied organisms not well represented in databases. Commercially-available hybridisation kits will function less effectively in this case, as database information will form the basis for their design. One way around this problem is to produce custom hybridisation probes tailored to a specific community, as demonstrated by Stewart *et al*. (2010). This approach can drastically improve rRNA removal and facilitate metatranscriptomic analysis, and is discussed in detail in chapter 4.

No rRNA removal mechanism is perfect and all metatranscriptomic datasets will need to be processed carefully in order that sequences likely to be of ribosomal origin are separated from the putative mRNA reads. There are a number of examples of metatranscriptomic surveys from a variety of environments, but currently and compared to metagenomic studies, they are relatively few. Reported numbers of rRNA reads in metatranscriptomic datasets range between 37% and 90% of the total (Stewart *et al*., Gifford *et al.*, 2011, Shi *et al*. 2011). Removal of these rRNA reads is important as their presence in a dataset of putative mRNA reads can lead to misannotations when searching for protein homologs to the sequences (Tripp *et al*. 2011). This is a relatively simple bioinformatics task within the spectrum of metagenomic analysis, and given that one of the strengths of high throughput sequencing is the volume of data produced, removal of rRNA reads from a metatranscriptomic dataset still leaves large numbers of sequences forming a considerable source of information concerning the gene expression of a microbial community.

1.2.4 Summary: Advantages and limitations of culture-independent techniques.

Care must be taken when analysing a microbial community through the extraction of nucleic acids from an environmental sample. Both extraction of nucleic acids in the first place and also subsequent sample processing, and in particular PCR amplification, will introduce bias in any assessment of species abundance using a metagenomic approach (Engelbrektson *et al*. 2010). Sequencing errors from the 454 pyrosequencing platform have been found to lead to overestimations of diversity in terms of inflation of the numbers of detectable OTUs. Kunin *et al*. (2010) found that stringent quality filtering on sequence output was important for producing sensible, accurate results in the subsequent analysis, with the homopolymer error of the 454 chemistry being a major cause of the problems. Although Illumina sequencing exhibits a lower error rate (Loman *et al*., 2012), it does not eliminate them and proper quality control of sequence data is extremely important.

Modern methods of qPCR, pyrosequencing and Illumina sequencing were all developed in response to a growing recognition of the limitations of pre-existing laboratory techniques, affecting many branches of biology. These innovations have been beneficial in studying environmental microbiology, the human microbiome and for investigating the cells and tissues of the higher Eukarya. In addition to many positive features, all of these methods have limitations and absolute drawbacks of which it is important to be aware. It is not the case that any one of these techniques is inherently better or worse than any of the others but simply that care must be taken when designing an experiment to choose the methodology most appropriate for a particular task, ensuring that the desired observations will be captured and the intended questions answered.

1.3 Applications of molecular microbial ecology

Microorganisms as pure cultures in the laboratory are ultimately an extremely unnatural phenomenon, as cells in the environment will inevitably be in constant contact, and possibly co-operation, with dozens of other species. Understanding what a species does in its natural environment is perhaps only truly elucidated by studying it within the context of that environment, even if laboratory culture is a possibility.

Cloning, qPCR and high throughput sequencing have been used to study communities of microorganisms in natural environments and have revealed information which would not have been discovered through traditional microbiology. This includes community structures and seasonal dynamics of environmental microbiota and novel enzymes. Molecular microbial ecology has led to the elucidation of new concepts such as that of the “rare biosphere” and a deeper appreciation of microbial abundance.

Metagenomic ribosomal gene data can be clustered to Operational Taxonomic Units (OTUs) and if clustering is based on a sequence similarity of 97% then OTU numbers correspond closely to species numbers (Wooley *et al*. 2010). Ribosomal genes are not perfect OTU markers but are frequently used, convenient and for most purposes are sufficiently accurate.

1.3.1 Molecular analysis of microbial communities

A large number of studies employing molecular methods and the examples that will be discussed here have been chosen either either due to their notability, as major publications highly cited in the field, or direct relevance to the results presented in later chapters. Warnecke *et al*. (2007) applied a functional

metagenomic approach to studying the termite hindgut. The contents of the third proctodeal segment of the hindgut, a microlitre-sized compartment containing a high density of microbes, was sampled from three worker termites of the wood-feeding *Nasutitermes* genus. The DNA extractions were used to produce a shotgun clone library. Sanger sequencing of the cloned sequences revealed a huge diversity of enzymes with glycoside hydrolase functions. Glycoside hydrolase enzymes would be expected to be numerous in this environment where symbiotic microorganisms assist the host in the digestion of the cellulose-rich wood diet the insects subsist on. A total of 700 glycoside hydrolase genes were found, from 45 different glycoside hydrolase families. Phylogenetic analysis revealed that some of these appeared to constitute novel clusters of proteins.

Analysis of a clone library produced by amplification of eukaryotic ribosomal SSU genes from DNA extracted from anoxic sediments identified novel lineages of eukaryotes at the kingdom level (Dawson & Pace 2002). This demonstrates the potential power of culture-independent techniques when applied to communities of organisms that have not been studied extensively due to the difficulty of growing them under laboratory conditions; there is scope for a great amount of diversity to be unearthed. Luo *et al*. (2005) utilised a clone library approach to study another anaerobic sediment at a spring where the waters were themselves hypoxic and contained high levels of dissolved sulphides. No Kingdom level diversity was identified in this study but fungi were found to be an important component of the community with a diverse range of fungal lineages represented by the clone library, including some sequences which appeared to constitute novel phyla.

McDonald *et al*. (2008) used qPCR to study the prevalence of the *Fibrobacter* genus in landfill leachate. *Fibrobacter spp*. have only ever been isolated from the herbivore intestinal tract, and primarily the rumen, but *Fibrobacter*-specific primer sets were used to amplify 16S rRNA genes from landfill sites. Application of qPCR showed that *Fibrobacter* could be detected in landfill leachate samples and in one case, *Fibrobacter* was shown to represent > 10% of the total population, detected in slightly higher quantities than in two rumen samples. A combination of nested PCR detection, cloning and qPCR revealed that Fibrobacter *spp*. can be detected in freshwater lakes and it would appear that novel lineages exist in these environments as cloned sequences cluster together distinctively and differently to those from other sources (McDonald *et al*., 2009).

Similar cloning approaches revealed a role for cellulose degradation in freshwater ecosystems for the *Micromonospora* genus. Cellulose-colonising biofilms in lake water were used as the source from which to isolate strains of

Micromonospora and some of these isolates were shown to be highly efficient cellulose degraders (De Menezes *et al*., 2008). Sequencing and phylogenetic analysis of the *gyrB* gene from these isolates demonstrated two specific clusters of *Micromonospora* isolates from the lake environment, one group of highly active metabolisers of cellulose clustering closely with the known cellulose degrader *M. chalcea*, and the second group forming a cluster which represented a probable novel lineage of the genus.

1.3.2 High-throughput sequencing of Marine Environments: towards an understanding of population structure and behaviour.

Early on in the history of high-throughput sequencing, the 454 pyrosequencing technology was enthusiastically adopted by research groups interested in microbial community structures of marine environments. High-throughput sequencing is especially suitable for studying this environment where a few species predominate but a large amount of low-abundance diversity exists. Many of the high-throughput sequencing studies based on either amplicon sequencing or on a shotgun metagenomic approach have been carried out in these environments.

One of the first examples was published by Sogin *et al*. (2006). They used PCR to amplify the V6 hypervariable region of the 16S rRNA gene of several environmental samples. These amplicon libraries were sequenced on the 454 genome sequencer platform and generated approximately 118,000 reads, which at the time was an impressive quantity of data. Diversity estimates on the communities sampled were calculated to be an order of magnitude higher than any other study of a microbial community previously published at the time. Clustering and rarefaction analysis of OTUs at a similarity level of 97% revealed that the community sampling achieved was far from comprehensive and greater sampling, and more sequencing, were both highly likely to increase the number of unique sequences and OTU clusters detected. This study was very much in the early days of high throughput sequencing. Sogin *et al*. found it necessary to limit their analysis to the first 100 bp of the reads produced as pyrosequencing technology at the time exhibited a severe drop in quality for bases beyond 100 bp. Limitations notwithstanding, the point was well made that there was a great deal of microbial diversity present in ocean waters. The findings were among the first to follow a pattern which has now become a familiar theme of environmental DNA sequencing projects, which is that the microbial communities in the ocean (and most other environments) appeared to be dominated by large populations of one or a few particular groups of organisms but that a long

tail existed of a large number of less numerous OTUs which contributed a greater part of the overall diversity of the community.

This early study revealed some interesting details about the microbiota of the oceans and firmly established the presence of a rare biosphere, that is, a large amount of microbial diversity which exists at very low abundance compared to a handful of dominant species; but it also posed many questions. Unsurprisingly, the microbiota of different sites were found to have different compositions. Specific groups, highly abundant at one location, could be a minor constituent at another. It was not clear, from samples taken as a single time point, whether minority members of the population had the potential to increase in number under the right conditions. Questions about ecology and functional role of the rare biosphere were still unanswered.

Building on this early work, later studies were able to take advantage of improvements in technology and in the affordability of high-throughput sequencing. A later investigation of the rare biosphere of the oceans sought to understand its distribution and behaviour, using a similar approach of V6 hypervariable region amplicon tags, by comparing the rare and abundant constituents of the population at several sampling points (Galand *et al*. 2009). This was put to test the hypothesis that the rare biosphere has a cosmopolitan distribution in the oceans, being constantly distributed by ocean movement and being protected from loss to lysis from bacteriophage infections and predation due to their low numbers. However, the rare biosphere did appear to have a biogeography, to be limited to certain areas and to be adapted to those areas. Similarity of the microbiota was greater for both rare and abundant phylotypes between samples that were from environments that were more alike in physical and chemical terms (e.g. deep water) than for samples that were closer together geographically. Therefore, the rare phylotypes constitute members of the microbial community specifically adapted to the environment in which they are found and may have important roles in the marine ecosystem.

Andersson *et al*. (2010) used Pyrosequencing to study temporal shifts in the abundance of bacterioplankton throughout the year in the waters of the Baltic Sea. They were able to show that the population was in a state of flux, exhibiting a high level of seasonal variation throughout the year. The sequencing of DNA from various sampling points yielded a total of 4624 OTUs from various times of the year. Of these, 1182 OTUs appeared in only a single seasonal sample and only 76 were observed at every sampling point. The data also suggested an annually recurring pattern as two samples taken twelve months apart in May were very similar. One

OTU was found to represent less than 0.01% of the reads in a sample taken in July, but 10% of the reads in a sample taken the following May.

This analysis demonstrates above all a high level of seasonal variation in the relative abundance of members of the population, and reveals that members of the rare biosphere can become highly abundant at certain times of the year, under the right conditions. What constitutes the “rare biosphere” might then depend on the time of year that sampling was carried out. This variation was found in samples taken at 3 m depth, and deeper samples, less exposed to the direct effect of the sun and seasonal factors, might exhibit far less variation, but the capacity of members of the rare biosphere to respond to changing conditions and become dominant members of the microbiota was clearly established. Low abundance does not equate with a lack of importance in ecological terms.

High-throughput sequencing projects have furthered knowledge of marine microbial ecology and provided a detailed and comprehensive overview of the communities that inhabit these environments. In this way, communities which comprise thousands of species can be monitored over time and understood. Metagenomics remains, however, somewhat descriptive, even if effort is made to produce data that provide a highly accurate quantitative assessment of the abundance of the member of a community. A functional approach to metagenomic sequencing can add an extra layer of utility to metagenomic datasets.

1.3.3 Functional metagenomics.

To fully understand a microbial community it is necessary to address both *who are they?* and, *what are they doing?* Since the concept of the “rare biosphere” was first floated (Sogin *et al*., 2006), addressing the question of *who?* has become routine. A simple metagenomic inventory of organisms present in a specific ecosystem is now a relatively simple objective to achieve. Going one step further and discovering exactly what the roles of the inhabitants of an ecosystem are, is a great deal more challenging. Additionally, environmental microbes harbour a large amount of genetic diversity and within their genomes there is a great untapped potential for the discovery of novel functions, for example exploitable enzymes. Several studies have employed high-throughput sequencing in an attempt to examine the genes present in a metagenome, the genes expressed in an environmental sample through metatranscriptomics, and also clone libraries to screen for novel genes.

Functional metagenomics has been used to investigate microorganism-mediated plant biomass degradation. There is commercial interest in discovering

new enzymes for use in processing plant biomass into biofuels, as existing processes are limited by the inefficiency of currently available biocatalysts (Alonso *et al*. 2010). High throughput sequencing techniques can be employed to carry out an in-depth survey of a metagenome from an environmental sample. These surveys can be used to produce catalogues of genes from uncultivated micro-organisms which can be functionally annotated to identify protein encoding genes (De Filippo *et al*. 2012; Schnoor, 2011) and this can be harnessed for the discovery of novel enzymes which may have better substrate specificity, high specific activity or better thermal and chemical tolerance than those previously known. Metagenomic screening for these enzymes is backed up in the context of this thesis by the specialised CAZy (Carbohydrate-Active Enzymes) database which documents the various families of enzymes known to have activities against complex polymeric carbohydrate compounds such as glycoside hydrolases (Cantarel *et al*. 2009).

454 pyrosequencing was used to produce a catalogue of the metabolic potential of fibre-adherent microbiomes of three different bovines by assessing the number of sequences corresponding to families of glycoside hydrolase and cellulosome proteins (Brulc *et al*. 2009). A comparison was made between the fibre-adherent metagenomes and a pooled metagenome sequenced from the liquid rumen contents of all three animals studied; the pooled metagenome did in fact reveal the same numbers of glycoside hydrolase family members as any of the individual fibre-adherent metagenomes. Colonisation of fibre in the rumen appeared to be spearheaded by organisms that attacked accessible side chains of plant polysaccharides, with a second subset of organisms equipped to degrade recalcitrant cellulose and hemicellulose polymers replacing the initial colonisers at a later stage.

Hess *et al*. (2011) used a similar approach to study the metagenomes associated with nylon bags containing switchgrass inserted into the rumen of a fistulated cow. Paired-end Illumina sequencing of several metagenomic libraries produced a total of 1.5 billion read pairs. ORF prediction led to the identification of 2,547,270 potential protein coding sequences. ORFs were analysed by searching them against Hidden Markov Models of CAZy gene families. 27, 755 ORFs were found to be carbohydrate active genes, i.e. encoding glycoside hydrolase functionality, a carbohydrate binding domain or other related function. Many of these had a similarity to sequences in the NCBI NR database below 75%, implying a large number of novel genes were detected. PCR amplification of some of the novel genes was achieved from metagenomic DNA samples from the same source, yielding

products which could be cloned, their encoded proteins expressed, and activities confirmed with biochemical assays.

Hess *et al*. were also able to assemble some partial genomes from the metagenomic data. The completeness of these assemblies varied between 93% and 60%. The extremely high depth of the sequencing project carried out as part of this study was shown to be highly effective for discovery of novel full-length genes and for the assembly of draft genome sequences of uncultured organisms.

Functional metagenomic work can reveal a wealth of enzymes encoded in the genomes of microbes, which, sculpted by the selective force of millions of years of evolution, can be highly efficient carbohydrate-degrading mechanisms. Although high-throughput sequencing only generates *in silico* data, the information produced can still be immensely valuable for guiding mining of metagenomes, by PCR-amplification of gene sequences detected by sequencing efforts, and subsequent cloning of PCR products to produce whole, characterisable, usable proteins (Li *et al*. 2009)

1.3.4 Environmental metatranscriptomes.

Once again, early adopters of metatranscriptomics were marine studies. One of the first environmental metatranscriptomic studies to be undertaken was designed to investigate the gene expression in ocean surface waters (Frias-Lopez *et al*., 2008). This early study identified high level of expression of certain pathways but perhaps more tellingly many of the sequences were found to have no hit to the NCBI NR database or to constitute a hypothetical protein of unknown function.

McCarren *et al*. (2010) used metatranscriptomics to study the microbial processes which effect turnover of dissolved organic matter in ocean surface waters. Seawater microcosms were set up and dissolved organic matter was added at intervals; periodic sampling of the microcosms was carried out to monitor the response of the microbial community to the dissolved carbon source. This study was able to demonstrate taxon-specific responses to dissolved organic matter, and specific up-regulation of certain metabolic pathways, some related to utilisation of organic matter but others involved in two-component signalling and chemotaxis.

In addition to using metatranscriptomics to survey gene expression in communities, some studies have been recently used as a mechanism for discovery of novel genes, which could be termed a functional metatranscriptomic approach. Lehembre *et al*. (2013) used metatranscriptomics to study eukaryotic genes responsible for conferring heavy metal resistance in contaminated soils. Their sampling process involved total RNA extraction and selection of polyA mRNAs.

These mRNAs were converted to cDNA and cloned and expressed for screening of resistance conferred on the transformants. This approach resulted in the discovery of novel genes, with no database homologues, encoding for cadmium resistance, demonstrating that sampling of a metatranscriptome can lead to the discovery of novel genes. The use of metatranscriptomic sequencing has also been utilised for the purposes of gene discovery, as described by Takasaki *et al*. (2013). Here, soil samples were enriched with cellulose and the eukaryotic mRNA sequences from the soil sample were purified and sequenced. Sequence reads which were putative protein coding sequences were compared to the NCBI NR database to identify those containing glycoside hydrolase functionality with a probable role in cellulose breakdown. Although this identification was based on database homology, which precludes the discovery of truly novel proteins, the low sequence similarity of some of the database hits suggested that the proteins from the metatranscriptome could represent enzymes with known activities, but new specificities or mechanisms.

1.3.5 Analysis challenges posed by large datasets.

Pyrosequencing has always generated longer reads than other high-throughput platforms but the technology appears to have hit an overall upper limit in terms of total data output. Illumina platforms produce huge amounts of data now but this may be in itself a limitation. Millions of reads, constituting datasets containing several gigabytes of information, requires powerful computational resolution. Short read aligners have been developed that can very efficiently align millions of reads to a reference sequence. Sequencing the genome of an organism for which there is a reference sequence is not a major challenge. Sequencing the genome of an organism for which no reference sequence exists is also very feasible as a large number of assemblers have been developed for *de novo* assembly, although these programs are resource demanding and need powerful computers to run. Metagenomics, and metatranscriptomics, though, create major further complications to this process.

High throughput sequencing data is described in terms of *depth* and *coverage*. Depth refers to the absolute number of reads generated by a sequencing run. Coverage refers to how many times each base is “covered” by a read. When working with a single organism and sequencing only one genome, it is possible to sequence the genome in its entirety multiple times and achieve a coverage of 5 to 10-fold. Higher levels of coverage facilitate assembly, or allow variations in genomes to be called with high confidence (e.g. SNP calling). Coverage is a relative term; the same depth of sequencing obtained from different samples will generate different

levels of coverage. Metagenomics, and metatranscriptomics, suffer in that even with great depth the coverage is usually fractional unless the microbial community being sequenced is unusually simple. Fractional datasets are extremely difficult, if not impossible, to assemble. Hess *et al*. (2011) did demonstrate that assembly of draft genome sequences was possible from metagenomic data but the assemblies will not be complete, and this also requires a massive sequencing effort to generate the level of coverage needed; something that is prohibitively expensive for many research groups.

Approaches other than sequence assembly are needed to deal with large metagenomic datasets. Previously, searching against blast databases has been heavily employed. The BLAST algorithm is not well-suited to very large datasets however and is particularly unsuited to aligning large numbers of reads from a shotgun metagenomics run to a large database such as the NCBI non-redundant protein database. The MEGAN software was developed for use with metagenomic data and provides a very effective visualisation of community diversity and functional content (Huson *et al*., 2012; 2011). However, it requires an output file from a BLASTx search as input to generate summary data. Pyrosequencing produced datasets of hundreds of thousands of sequences and using BLAST to analyse this sort of output was feasible. BLAST cannot be so sensibly applied to the millions of reads generated by Illumina sequencing.

MG-RAST is a web service that has been used to analyse hundreds of metagenomes and metatranscriptomes. It utilises more efficient search algorithms than the BLAST and provides a breakdown that is both phylogenetic and functional. MG-RAST is becoming less suitable as datasets grow larger, and standalone software becomes more desirable than attempting to transfer large amounts of data to a webserver. It is also somewhat inflexible as a web service, since it is impossible to customise or tailor the analysis pipeline in any way or combine it with other software easily.

Short read aligners such as Bowtie (Langmead *et al*., 2009), Bowtie2 (Langmead & Salzberg, 2012) and BWA (Li & Durbin, 2009) have been developed specifically to align short reads to reference sequences. They are capable of handling millions of short reads and the programs can run at high speed with modest computer hardware. The high efficiency of these programs makes them potentially very useful for dealing with large numbers of unassembled short reads in a metagenome or metatranscriptome, and they are beginning to be used in this area.

Metabin (Sharma *et al*., 2012) and Genometa (Davenport *et al*., 2012) were developed to combat this exact problem. Genometa is a fork of the Integrated

Genome Browser (IGB) software and it makes use of the bowtie aligner to classify short metagenomic reads to a database of reference genomes, generating a histogram that visualises the distribution of the reads. The software provides many other customisable view options and data export functions. Metabin efficiently characterises large numbers of short reads using the BLAT (Blast-like Alignment Tool) algorithm (Kent, 2002) as a faster alternative to BLAST. Metabin was shown to be a fast and accurate classifier of metagenomic short read data and is also capable of outputting data in a format that can be used to generate taxonomic trees, and pie charts displaying proportions of the populations assigned to various taxonomic groups, and the Metabin webserver allows multiple Metabin output files to be compared.

It is becoming feasible to use shotgun metagenomic analysis to investigate microbial community structure and to avoid SSU rRNA gene amplicon-based surveys. There is a major limitation on the analysis provided by both Metabin and Genometa, however, which is that both of these programs still focus almost entirely on taxonomic classification and still provide only limited functionality towards discovering and characterising novel genes of interest with datasets. They are a useful first step, but gene discovery still requires a tailored approach and it is likely to be some time before a single program becomes available that can perform this task.

1.4 Molecular microbial ecology and the metabolism of cellulose by anaerobic microorganisms.

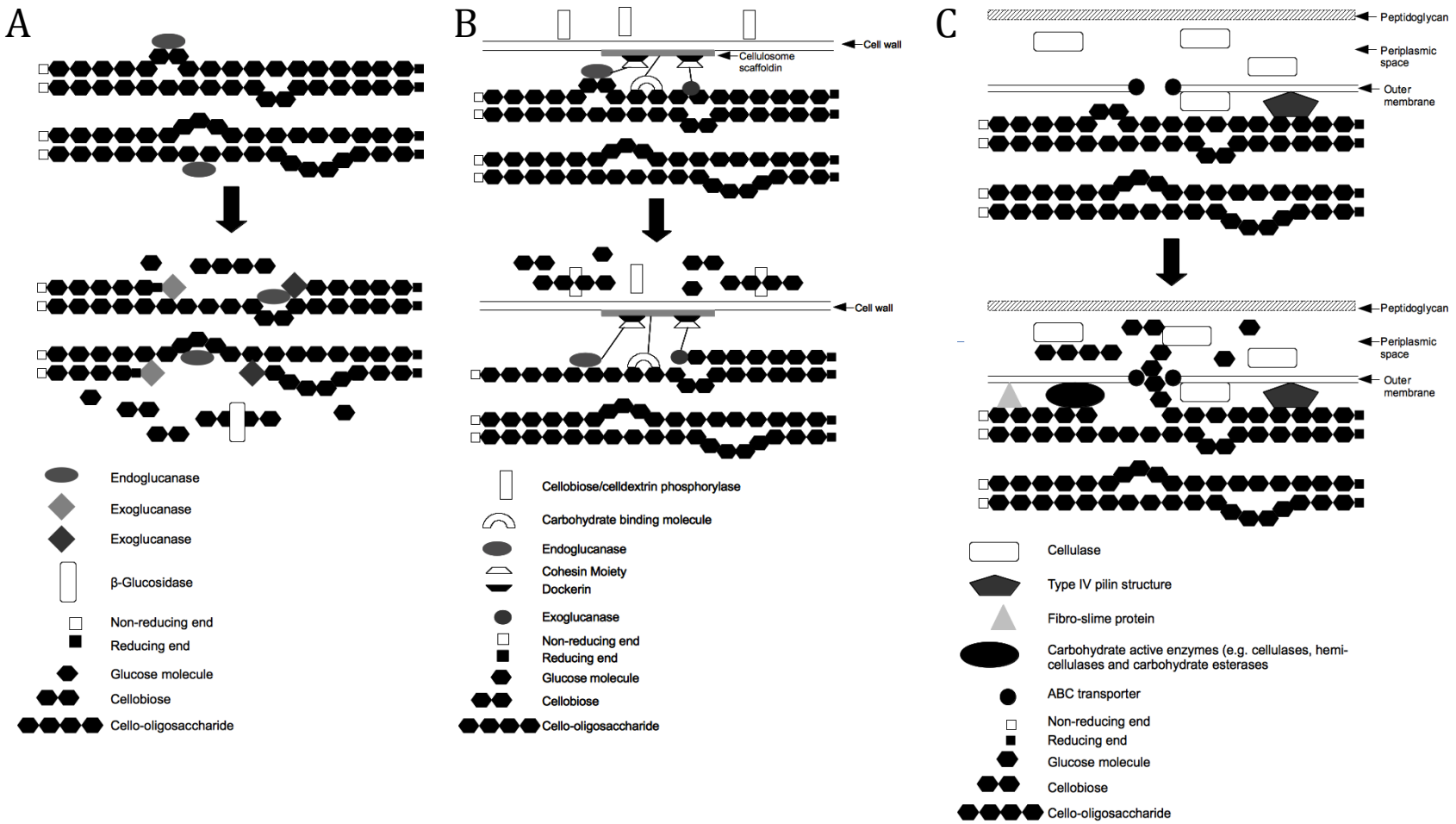
1.4.1 Cellulose and microbe-mediated cellulose breakdown.

Cellulose is a glucose polymer and an extremely stable compound found as a major structural component in plants (Beguin & Aubert, 1994). It is therefore highly ubiquitous in the environment and an abundant carbon source, but it is also highly recalcitrant; the properties that make it an excellent structural material for plants also make it resistant to enzymatic digestion (Lynd *et al*. 2002). The huge quantity of cellulosic material globally has led to an interest in using it as a feedstock for biofuel production although a cost-effective industrial scale cellulosic biofuel pipeline has remained elusive, mainly due to the high cost of enzymatic treatments needed to break down the polymer into fermentable material (Lynd *et al*. 2002). Microorganisms have been utilising cellulose as a carbon source for millions of years and studying the species responsible for cellulose breakdown may illuminate industrially applicable enzymes and organisms for use in the biofuel industry.

Historically, the majority of the work undertaken to study the enzymatic machinery used by microorganisms has focussed on aerobic fungi. *Trichoderma reesei* in particular is a heavily studied aerobic cellulolytic fungus and has been the principle source of industrial cellulases (Martinez *et al*. 2008). More recently attention has swivelled towards anaerobic cellulose degradation, which is carried out by bacteria and fungi and represents approximately 10% of worldwide cellulose breakdown (Leschine, 1995).

Studying anaerobic cellulolytic bacteria has yielded some new mechanisms of enzymatic attack against cellulose. Cellulolytic clostridia use a large, extracellular complex of proteins called a cellulosome to attack polymers of cellulose (Fontes & Gilbert 2010). Cellulosomes are assembled from enzymes which provide catalytic activity, and structural proteins termed scaffoldins. Scaffoldins and the cellulosome enzymes possess recognition modules termed cohesins and dockerins respectively which allow high-affinity recognition between the structural and functional components of the cellulosome and thereby mediate assembly of the structure (Bomble *et al*. 2011). Other anaerobic bacteria appear to utilise a system different from both the clostridial cellulosome and the free secreted cellulase system known from aerobic fungi, where individual cellulose polymers are removed from plant matter and internalised into the periplasmic space (Wilson, 2009). *Fibrobacter succinogenes* and *Cytophaga hutchinsonii* have been associated with this proposed mechanism in particular which as yet remains a hypothetical mechanism. Cellulolytic fungi have been found to possess proteins with dockerin-like sequences and seem to have their own, independently evolved, version of cellulosomes (Fontes & Gilbert 2010). As so much remains to be understood about exactly how anaerobic microbes breakdown cellulosic material, anaerobic cellulose degrading communities might harbour a great deal unknown novelty and are an obvious target for searching for new genes.

Figure 1.1 presents a schematic of currently understood and proposed mechanisms by which microorganisms breakdown cellulosic material using enzymatic machinery.



Fugure 1.1: Legend on following page

Figure 1.1 (Previous page), reproduced from Ransom-Jones *et al*. (2013). (a) The free cellulase system, utilised by T. reesei and other fungi responsible for carrying out aerobic degradation of cellulose, relies on the secretion of a range of enzymes whose synergistsic activity is able to attack the polymer. (b) The cellulosome system, used by anaerobic clostridia and possibly anaerobic fungi too, involves the assembly of a complex of enzymes on the cell wall. The cellulosome incorporates carboyhydrate binding domains, which maintain close proximity to the substrate, as well as maximising the synergy of a range of glysocide hydrolase enzymes. (c) The proposed mechanism of *Fibrobacter succinogenes* and *Cytophaga hutchinsonii* is yet to be confirmed but appears to involve the disruption of cellulose polymer chains by secreted or outer membrane enzymes and subsequent internalisation of individual cellulose chains for further break down.

1.4.2 Molecular approaches to studying and isolating cellulolytic enzymes.

Large-scale metagenomic screening has been carried out before in environments where cellulases are expressed by microorganisms that have adapted to utilise this highly abundant but also extremely recalcitrant carbon source. The termite hindgut has been investigated (Warnecke *et al.*, 2007) as has the bovine rumen (Brulc *et al.*, 2009; Hess *et al.*, 2011). Beyond the digestive tract of animals whose diet contains a high proportion of cellulosic material, marine microbial communities (Edwards *et al.*, 2010) and communities resident in a bioreactor (Krause *et al.*, 2008; Krober *et al.*, 2009; Schluter *et al.*, 2008) have also been studied.

Cloning has been used to identify, express and functionally characterise cellulases from environmental samples. The issue with utilising cloning in this manner is that the hit rate is extremely low and thousands of clones may need to be screened to identify a single functional carbohydrate active protein. A fosmid library from a bovine rumen metagenome yielded two highly active cellulases, as determined by screening on CMC-containing agar representing a hit rate of 1 in 476 (Rashamuse *et al*. 2013). The enzymes were novel proteins, however, with similarities of no greater than 75% to previously isolated sequences. A library produced from a soil metagenome consisting of 3024 Bacterial Artificial Chromosomes (BACs) yielded one positive clone, found to contain an endo-β-1,4-glucanase that had a closest match of 39% similarity to any other protein in a database (Liu *et al*. 2011). These cloned genes and the proteins they encode have revealed useful traits such as a tolerance and preservation of catalytic activity in a range of temperature and pH conditions. Properties such as these are helpful when screening for enzymes with potential industrial applications.

Cloning studies have produced interesting proteins but also indicate that there is a great deal of genes of unknown function within the genomes of environmental bacteria that are yet to be discovered. Many of these species are likely to be difficult to cultivate in a laboratory and a culture-independent approach is the best way of identifying them and elucidating their metabolic potential and arrays of enzymes.

1.5 Sampling sites and experimental design for studying environmental microorganism-mediated cellulose breakdown.

Two different but similar environments were studied here. They were chosen as known anaerobic environments, where cellulose is present and is broken down

under anoxic conditions by communities of microorganisms. Previous studies have chosen to identify novel cellulases and other carbohydrate active enzyme categories by studying environments, and particular niches within those environments, where cellulose degrading organisms are known to be present in high numbers. Examples of this include metagenomic studies of organisms adherent to lignocellulosic fibre and plant matter within the bovine rumen (Brulc *et al*. 2009, Hess *et al*. 2011). Focussing on the rumen contents that were likely to be colonised by lignocellulose degrading organisms does seem to have resulted in the detection of a large number of carbohydrate active genes. Beyond this example, use of cellulose baits has been used to recruit cellulose degrading organisms in environments where they are less numerous and has resulted in the detection of interesting organisms and genes (Edwards *et al*., 2010, de Menezes *et al*., 2012).

One of the environments studied was a disused landfill site at Bromborough Dock (Wirral, UK), which contained municipal solid waste (MSW) and had been capped at the time of sampling. The leachate within the landfill site contains dissolved and suspended waste matter, is accessible via a system of “risers” which allow the liquid to be pumped to the surface, enabling sampling. MSW in landfill will always contain a sizable component of cellulosic and hemi-cellulosic material. Quantities vary between sites, but the amount of cellulosic waste in a landfill site varies between 20 and 50% (Barlaz, 2006). There is, therefore, a great deal of carbon available for organisms able to access it in the form of cellulosic material.

Bromborough Dock has previously been used as a sampling site for molecular ecological work on the microbial communities in landfill leachate, and has been shown to harbour cellulolytic communities of microorganisms including fibrobacters, the detection of which appears to indicate novel members of the genus, related to but distinct from the characterised residents of the rumen known for their cellulose-degrading capacity (McDonald *et al.*, 2008; McDonald *et al.*, 2010).

The freshwater lake Esthwaite Water in the Lake District national park was selected here as a site for metatranscriptomic sampling. Esthwaite water (Fig 1.2) is situated in a fertile valley and is one of the more productive lakes in the Lake district (George *et al.*, 2000). The average depth of the lake is 6.4m although the samples for this study were derived from the deepest part of the lake, which is approximately 14m in depth subject to seasonal variation, where conditions in the lower part of the water column and the sediment are anoxic. Esthwaite water has been used previously as a sampling site for various studies focussing on cellulolysis, including investigations into the molecular ecology of *Fibrobacter* spp. (McDonald *et al.*, 2009) and *Micromonospora* spp. (de Menezes *et al.*, 2008; 2012).

Fig 1.2 Esthwaite Water in the Lake District.

1.6 Aims of the project.

Firstly, work previously undertaken on the landfill leachate of Bromborough Dock landfill site was continued. Data had been collected revealing a presence of cellulolytic lineages of bacteria in the leachate of the landfill site previously. Samples of DNA extracted from the landfill leachate were used for a qPCR analysis to determine the abundance of these lineages in the community, in order to confirm whether they were numerous and major contributors to the population. A qPCR analysis was also performed on cellulose baits incubated in microcosms containing leachate harvested from the landfill site to assess the colonising community and determine if any of the cellulolytic lineages apparently present in the leachate might be enriched in the biofilms covering the cellulosic cotton bait material. Previous published work had established a presence of unknown *Fibrobacter spp*. which would if present almost certainly contribute to cellulose breakdown (McDonald *et al*. 2008) and many studies in the past have implicated clostridia as mediating cellulose breakdown in landfill (Westlake *et al*., 1995; Burrell *et al*., 2004). Using qPCR, the specific presence and abundance of fibrobacters and clostridia on cellulose baits could be assessed and compared with the abundance in leachate generally in an attempt to confirm that they are associated with, and by extension implicated in the breakdown of, cellulosic material in landfill sites.

Secondly, a greater part of the work focuses on studying the anaerobic microbial community in the sediment of the lake Esthwaite Water with a view to assessing the phylogenetic and functional content of the community and discovering expressed genes from this environment involved in cellulose degradation. Freshwater anaerobic communities have not been previously screened for expressed cellulolytic activity and there was potential for new discoveries. Using the pre-established mechanism of producing cellulose baits with de-waxed cotton, the sediment community was sampled by harvesting baits that had resided at the bottom of the lake and become colonised by microbes present in the sediment. Harvested baits were used for nucleic acid extraction and subsequent sequencing. It was hypothesised that the cellulose baits would become associated with a biofilm enriched with cellulolytic organisms and the transcripts for gene involved in cellulose degradation would be present in the database in elevated numbers with respect to a dataset based on an extraction just from the lake sediment alone. The DNA material was also used to generate a metagenome of the cotton bait colonising community and a fosmid library of high molecular weight DNA fragments was also generated, to serve as a resource for functional screening of the community metagenome. This multi-pronged approach was intended to provide an indication of the active genes

which could in turn inform analysis of the fosmid library, by screening for genes known to be present in the datasets obtained from sequencing.

Chapter 2: General Methods

2.1.Environmental sampling.

The samples from which environmental nucleic acids were extracted for use in this study were derived from a freshwater lake sediment, and from landfill leachate. Both the lake sediment and landfill leachate were sampled using cellulose baits. In addition, landfill leachate samples were investigated directly.

2.1.1Production of cellulose baits.

Cellulose baits were produced according to a method adapted from Wood (1988). The method was carried out in a simplified form here as Wood describes a protocol for the preparation of highly pure cotton for quantitative chemical assays, whereas this study only required a cellulose-rich substrate that would favour colonisation by cellulolytic microorganisms. Briefly, cotton string was extracted in a soxhlet apparatus with chloroform for 18 h, left under a fume hood to evaporate the solvent and then washed again with ethanol for 18h. The cotton was then boiled in 1% NaOH for 4 h, and washed in warm water for a further 4h and soaked in cold water overnight. De-waxed cotton was placed into nylon mesh bags (Fig. 2.1) before being placed *in situ* for microbial colonisation. Nylon mesh acted as an inert material to hold the string in place while leaving it available for colonisation by the local microbiome.

2.1.2 Landfill Leachate sample collection.

Landfill Leachate total community DNA samples were available from previous work. Samples were previously collected from landfill sites across Northwest England in October 2005 from Bidston Moss, Risley and Bromborough Dock municipal waste landfill sites as described in McDonald *et al*. (2008). Samples of leachate were collected from the Bromborough Dock site in March 2007. One 10L volume of leachate comprised a combination of leachate of risers 3 and 4 (due to problems with the pumping system it was impossible to sample these areas independently) and another 10L volume from riser 5, collected in carboys (Figs 2.1 and 2.2) These were maintained at room temperature in the laboratory and used as microcosms for studying colonisation of the cellulose baits that were suspended in these carboys, and left for a period of one month before harvesting for nucleic acid extraction.

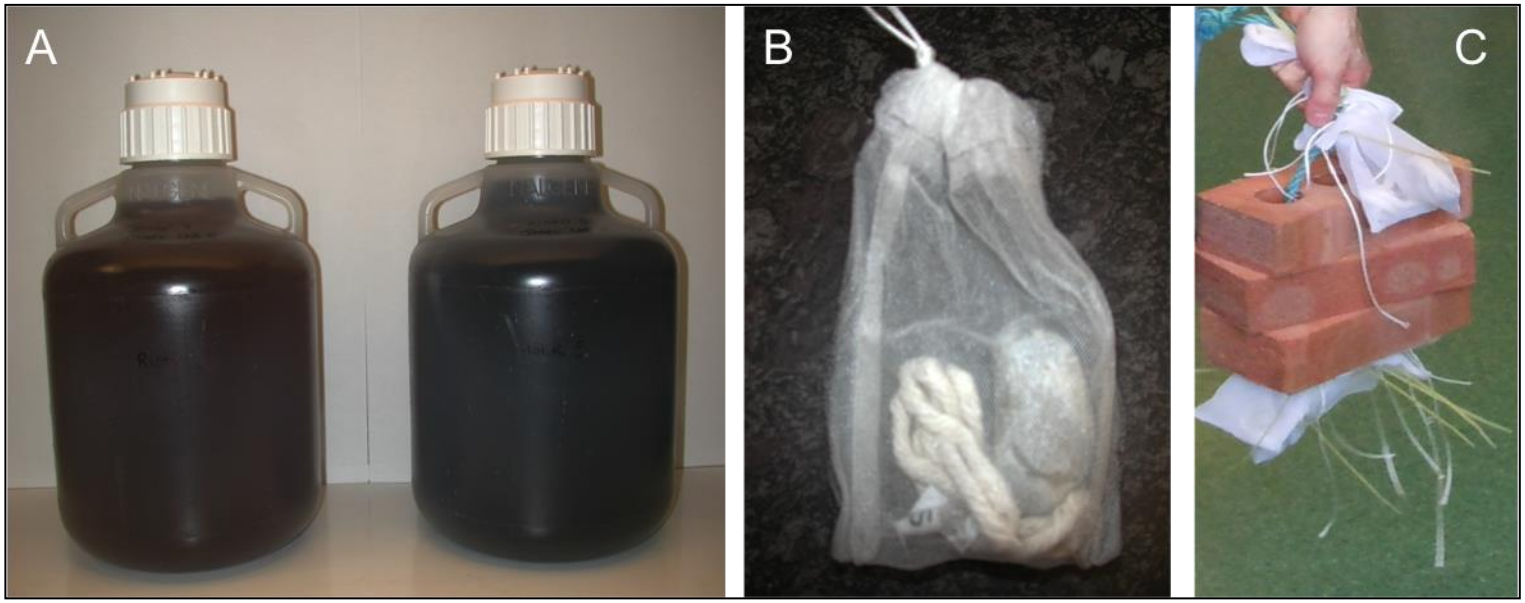


Figure 2.1: Sampling equipment and microcosms. Landfill leachate was collected in carboys (A) and stored in the laboratory. Cotton string in nylon bags (B) was introduced into the microcosms and incubated in order that the colonising biofilm be used as a source of nucleic acid. Nylon bags containing cotton string were weighted with bricks (C) in order to allow to the cotton to be maintained in the anaerobic lake sediment to allow colonisation of the cotton by a biofilm from this environment and subsequent analysis of the biofilm community.

.

Figure 2.2: A “Riser” sampling point at Bromborough Dock.

2.1.3. Lake Sediment Sampling.

Cellulose baits were attached to anchors that ensured the baits would be weighted down into the anoxic lake sediment. The initial sampling campaign ran from August – October 2010. Baits were introduced on 12/08/10 and one set was harvested after two weeks. Two further sets of baits were retrieved after six and eight weeks. After baits were drawn up from the sediment, they were immediately placed onto dry ice for transportation back to the laboratory. Samples were either processed for nucleic acid extraction on the same day, or stored at -80^{o}C for processing at a later date.

A second sampling campaign was carried out, running from July – October 2012. Cellulose baits were retrieved just once in October after approximately 10 weeks, and transported on dry ice as before. Baits were stored at -80^{o}C until use.

2.2. Nucleic Acid Extraction.

2.2.1Preparation of RNase-free solutions and equipment.

Solutions used in the extraction of both RNA and DNA from cellulose baits were treated with DEPC in order to inactivate RNAses and DNAses and protect the samples from enzymatic damage. Solutions prepared in glassware were treated with the addition of DEPC to 0.05% v/v and then incubated overnight at room temperature, with rotation, and finally autoclaved to destroy the remaining DEPC. Plasticware was autoclaved before use in containers washed with RNase Zap solution (Ambion) and rinsed with DEPC treated water. Pipette exteriors and gloves were treated liberally with RNase Zap before every experiment where RNA was to be handled.

2.2.2. Nucleic acid extraction using the “Griffiths method”.

Nucleic acids were extracted from samples using the method of Griffiths *et al.* (2000), which comprises mechanical bead-beating and chemical extraction, and was carried out briefly as follows. 5% (w/v) CTAB lysing buffer was prepared by combining equal quanitities of a 10% (w/v) CTAB in 0.7 M NaCl solution and 240 mM potassium phosphate buffer (pH 8). Bead-beating tubes were prepared by the addition of acid-washed glass beads (Sigma) to 2 ml screw-cap tubes. 0.5 ml 5% CTAB lysing buffer and 0.5ml Phenol:chloroform:isoamyl alcohol (25:24:1) were added to each tube. Typically, approximately 0.5g of colonised cellulose bait was processed in one tube.

Samples were subjected to bead-beating for mechanical disruption of cellular material, either in a fastprep for 30 seconds at a speed of 5 m s^{-1} or, as an equivalent treatment, in a MoBio Powerlyzer at a setting of 3400 RPM for 60 seconds. After bead beating, the samples were centrifuged at 17,000 x g at 4^{o}C for 10 min. The upper aqueous phase was decanted to an equal volume of chloroform:isoamyl alcohol (24:1) and the tubes were then centrifuged as before. The upper layer was transferred to a new tube and nucleic acids precipitated with two volumes of PEG solution (30% (w/v) PEG 6000 in 1.6 M NaCl) and left to precipitate for either two hours at room temperature or overnight at 4^{o}C. The nucleic acids were then pelleted by centrifugation at 17 000 x g for 10 minutes, and washed with ice-cold 70% ethanol in DEPC water. Residual ethanol and moisture were removed by evaporation for approx. 2 h in a fume hood or by vacuum centrifugation for 10 min at 40 ^{o}C. Pellets were resuspended in 50µl DEPC-treated water and stored at -20^{o}C, or at -80^{o}C for long term storage.

2.2.3. Purification of extracted nucleic acids for RNA-free DNA and DNA-free RNA.

The Griffiths method co-extracts nucleic acid. In order to produce RNA-free DNA extracts, the samples were made up to 150µl with nuclease-free water, RNase A was added to 100 µg ml^{-1} before incubation at 37^{o}C for 15 min. DNA-free RNA was obtained by treating RNA samples with TURBO DNA-free (Ambion) according to the manufacturer's instructions. Briefly, 5 µl of 10x DNase buffer and 3 units of TURBO DNase were added to 50 µl of the sample and incubated at 37^{o}C for 30 min. An additional 2 units of TURBO DNase were then added, followed by a further 30 min incubation.

In order to purify samples after nuclease treatment, 200 µl phenol:chloroform:isoamyl alcohol (25:24:1) was added. The tube was briefly vortexed and centrifuged at 10,000 x g for five min. The aqueous layer was transferred to a fresh tube, and an equal volume of chloroform:isoamyl alcohol (24:1) was added. The tube was vortexed and centrifuged as before, and the aqueous layer was again transferred to a fresh tube. The nucleic acid was precipitated with two volumes of 100% ethanol and 1/4 volume of 10M ammonium acetate, incubated for 30 min. at -80^{o}C, and centrifuged at full speed in a microfuge for 20 min. The supernatant was discarded and the nucleic acid pellet was washed with 200 µl ice cold 70% ethanol and air-dried, or dried by vacuum centrifugation for 10 min at 40^{o}C. The pellet was resuspended in 50µl nuclease-free water and stored at -80^{o}C until use.

2.2.4.Quantification and quality control of nucleic acid preparations.

A combination of methods was used. Reliable quantification was obtained using the Qubit fluorometer platform (Invitrogen). NanoDrop spectrophotometer (Thermo Scientific) readings were taken to assess purity of samples, based on the ratios of the absorbance values at 230 and 260 nm (A_{230}/A_{260}) and at 280 and 260 nm ($A_{260/}$ A_{280}) for chemical and protein contamination respectively. Visualisation of nucleic acids was achieved by agarose gel electrophoresis in order to inspect samples for signs of degradation and enzymatic damage, as evidenced by poor band integrity.

2.2.5 Extraction of genomic DNA from bacterial cells.

Genomic DNA was extracted from bacterial cell cultures according to the method of Cheng *et al*. (2006). Briefly, 1 ml of an overnight culture was centrifuged at full speed in a microfuge for 2 min in a microfuge tube. The supernatant was discarded and the cells were washed with 400 µl of STE buffer consisting of 100mM NaCl, 10mM Tris/HCl and 1mM EDTA at pH 8. Pellets were resuspended in TE buffer, comprising 10mM Tris/HCl and 1mM EDTA at pH 8. 100 µl Phenol was added and the tube vortexed for 60s. The tube was centrifuged at 14, 000 rpm for 5 min and the aqueous phase decanted to a fesh tube. 40 µl TE buffer and 100 µl chloroform were added, the tube mixed by inversion. Centrifugation was carried out as before and the aqueous phase was decanted to a fresh tube, quantified and stored at -20^{o}C.

2.3. Agarose Gel electrophoresis.

Conventional gel electrophoresis was used for size-checking of nucleic acid extractions and PCR products routinely. Electrophoresis for this purpose was performed using 1x TAE buffer, in a 0.8-2% agarose gel, adjusted for optimal resolution depending on the size of the fragments being analysed at the time. Agarose gels were stained with ethidium bromide added to a concentration of 0.5 ug ml^{-1}. Running conditions were also adjusted depending on size fragments being analysed and size of the gel used, but typically a constant voltage of 100V was used, for a time period between 30 min to 1 h 30 min.

2.3.1 Extraction of DNA bands from agarose gels.

In order to purify DNA bands from agarose gels, either the Qiaquick gel extraction kit (Qiagen) or the Bioline Isolate kit (Bioline) were used. The purifications were carried out according to the instructions provided by the manufacturer.

2.3.2 Pulse-Field Gel Electrophoresis

Pulse-Field Gel Electrophoresis was used for separation of DNA of high molecular weight. PFGE was carried out using a 1% low melting point agarose gel in 0.5× TBE buffer in a CHEF-DR III system (Bio-Rad) with a switch time of 1–6 s, voltage of 6 V/cm, 120° angle and run time of 18h. Staining of PFGE gels was typically performed after electrophoresis, by washing the gel in a 0.5 ug ml^{-1} solution of ethidium bromide.

2.4 End Point PCR.

PCR was carried out using either SuperTaq DNA polymerase (HT biotech), Phusion high-fidelity DNA polymerase (Finnzymes) or Herculase II Fusion DNA Polymerase (Stratagene)

2.3.1 Super Taq.

Reactions were carried out in 50 µl volumes. The reaction mix was made up with the following components: 0.2 µM each primer, 0.2 mM each dNTP, 1 x SuperTaq Buffer, 0.5 mM $MgCl_2$, 1 x BSA, 1 unit polymerase made up to 50 µl with sdd$H_2$0. Typical PCR cycling conditions consisted of denaturation at 94^{o}C for 5 min, 35 cycles with a denaturation step at 94^{o}C for 1 min, annealing at the specific temperature for the primer set for 1 min, extension for 1 min at 72^{o}C and a final extension step at 72^{o}C for 10 min. Conditions were altered if necessary to improve amplification and specificity.

2.3.2 Phusion.

Reactions requiring high sensitivity and fidelity were performed using Phusion polymerase (Finnzymes), which has a low error rate and has proof reading activity. Reactions were carried out in 25 µl volumes. The reaction mix was made up as follows: 0.2 µM each primer, 0.2 mM each dNTP, 12.5 µl of 2 x Phusion HF buffer, 1 µl of Phusion polymerase, made up to 25 µl with sddH_2O. Typical PCR cycle conditions were denaturation at 98^{o}C for 45 s, 35 cycles of 98^{o}C for 10 s, 30 s at the appropriate annealing temperature for the primer set, 72^{o}C for 30 s and a final extension at 72^{o}C for 5 min. In some cases, phusion reactions were performed in a 50µl volume if yield was poor, or to boost yields obtained from target DNA of low quantity. In this case, the amounts given above were adjusted accordingly.

2.3.3 Herculase.

The Herculase II Fusion DNA Polymerase (Stratagene) was also used when a high-fidelity polymerase was required. Herculase reactions were carried out in 50

µl volumes. The reaction mix was made up as follows: 0.25 µM each primer 0.25 mM each dNTP, 10 µl of 5x Herculase II buffer, 1 µl of Herculase II polymerase made up to final volume with sddH_2O. Typical PCR cycle conditions were denaturation at 95^oC for 2 min, 35 cycles of 95^oC for 20 s, 20 s at the appropriate annealing temperature for the primer set, 72^oC for 2 min and a final extension at 72^oC for 3 min. Conditions were altered if necessary to improve primer efficiency and specificity.

2.4 DNA clean up protocol

To remove residual contamination from DNA preparations, or to clean up DNA after PCR and other enzymatic treatments, DNA was cleaned using a spin column protocol from either the Qiaquick PCR purification kit or the Bioline PCR and gel kit. The protocols were carried out according to the manufacturer's instructions, but in both cases elution of the DNA from the filter column was carried out with water instead of the elution buffer provided with the kits to avoid interference with downstream applications. Additionally, elution steps were performed twice, to increase the yield, with two elutions of 30 µl for the Qiagen kit and 10 µl for the Bioline kit.

2.5 Quantitative PCR.

2.5.1 Reverse Transcription of RNA samples.

In order to generate single-stranded cDNA for qPCR experiments on environmental RNA samples, reverse transcription was carried out using Bioscript reverse transcriptase (Bioline). Reactions were set up with 1 µg RNA template, 0.2 µg of random hexamer primers and DEPC-treated H_20 added to a final volume of 12 µl. This mix was incubated for five minutes at 70^oC and then chilled on ice. 1 µl 10mM dNTP mix, 4 µl 5x Bioscript reaction buffer, 10 units of RNasin Plus Ribonuclease inhibitor (Promega) and DEPC-treated H_20 to 19.5 µl were then added, vortexed, and finally 0.5µl BioScript was added. The mixture was incubated at 42^oC for 60 min, after which it was heated to 70^oC for 10 min to stop the reaction and then chilled on ice and stored until use at -20^oC. DNA produced in this way was suitable for use in amplification reactions without additional purification.

2.5.2 Assay conditions.

qPCR was performed in an Applied Biosystems StepOnePlus Real-Time PCR system in a 96 well plate format. The reagents used were either Qiagen Quantifast SYBR green PCR kit, or Thermo Scientific SYBR Green qPCR master

mix. Experiments that were to be directly compared were always performed with the same reagents. For both reagents, reactions were performed in 25 µl volumes, and made up with identical quantities of components. Master mixes were made up with components calculated per 25 µl reaction, utilising 2.5 µl each primer (for 0.25 µM), 6.5 µl of SYBR green master mix, 12.5 µl nuclease-free water for each 25 µl final volume. 24 µl of the final master mix and 1 µl of template DNA, either 50 ng of unknown DNA or DNA of known copy number per µl for standard curve production, were added to each well.

2.5.3 Production of standard curves for relative quantification.

Relative abundance was used to enumerate target organisms as a proportion of the total bacterial population present in a sample using universal and group specific primers, or to compare the same target in two differently treated samples. Standard curves were generated as described in Smith *et al*. (2006) for this purpose. Triplicate serial dilutions ranging from 3×10^8 – 3×10^3 gene copy numbers per µl were prepared for each target gene and for each primer set.

Gene copy numbers were calculated according to the following formula:

$$\frac{100 \times [\text{concentration of sample in g/µl}]}{[\text{length of template}] \times 660} \times 6.022 \times 10^{23}$$

where length of template represents either the length of the linearised plasmid or of the PCR fragment.

Triplicate repeats allowed occasional spurious reactions to be disregarded from the analysis, and mean Ct values for each standard curve dilution and each unknown sample could be obtained from duplicate or triplicate values. Ct values were plotted against log gene copy number to generate standard curves for each primer set used in the assay. Linear regression of the data generated equations with which Ct values from unknown samples could be converted to values in terms of gene copy number of the target gene. Copy numbers determined with specific primer sets were expressed as a fraction of the copy number reported by the universal primer set run in parallel. Copy numbers determined when comparing different samples with the same set of primers was carried out by expression the lower sample as a fraction of the higher sample.

In order to generate standard curves, template DNA consisting of 16S rRNA subunit gene sequences of the target *Clostridium* groups was prepared from strains of *E. coli* harbouring plasmids containing cloned insert 16S rRNA gene copies, provided by Dr James McDonald (University of Bangor). A strain containing *Fibrobacter* 16S rRNA subunit gene was not available and this standard was

prepared by PCR amplification of DNA extracted from a pure culture of *Fibrobacter succinogenes*, also provided by Dr James McDonald.

2.5.3.1 Production of DNA template for standard curve generation.

Plasmid template DNA containing the 16S rRNA gene of *Clostridium* groups was prepared from overnight cultures of the host *E. coli* strains in LB. 1.5ml of culture supernatant was centrifuged, the supernatant discarded and plasmids extracted using the Qiaprep spin miniprep kit (Qiagen). Plasmids were linearised by restriction enzyme digestion before use in amplification reactions. All plasmid digestions were performed in 50 µl reaction volumes, digested at 37 ^{o}C for 1 h and inspected by agarose gel electrophoresis. For the *Fibrobacter* group, a PCR product of the 16s rRNA gene was subjected to further rounds of amplification to produce plentiful material for use in qPCR reactions. Table 2.1 summarises the genes used as templates for standard curves

Table 2.1 Source of Standards for qPCR experiments.

Template DNA*	Source	Purification method
C. aldrichii, for group III clostridia	*E. coli* strain containing plasmid with cloned *C. aldrichii* 16SrRNA gene	Miniprep, digestion with *Nco*I
C. sporosphaeroides for Group IV clostridia	*E. coli* strain containing plasmid with cloned *C. sporosphaeroides* 16SrRNA gene	Miniprep, digestion with *Nde*I
C. lentocellum for group XIV clostridia	*E. coli* strain containing plasmid with cloned *C. lentocellum* 16SrRNA gene	Miniprep, digestion with *Nde*I
F. succinogenes for *Fibrobacter* genus	PCR amplification of *F. succinogenes* 16S rRNA gene.	Size selection through electrophoresis and gel extraction.

*Clostridium groups as defined by Collins *et al*. 1994; *Fibrobacter* and *F. succinogenes* as described in Ransom-Jones *et al*. 2012

2.6 PCR primers.

Several PCR primers sets were used in this study, obtained from information in previous publications, designed specifically for particular purposes, or provided with specific kits. The primers used, their purpose and general characteristics are described in Table 2.2. The specific conditions are discussed in the appropriate sections.

Primer Set	Purpose	Sequence 5'-3'	Target	Annealing temp.	Amplicon Size (bp)	Reference
pA	End point PCR/Routine amplification	AGAGTTTGATCCTGGCTCAG	General Bacteria	55 °C	~1530	Edwards *et al*., (1989)
pH		AAGGAGGTGATCCAGCCGCA				
Chis 150	End point PCR/Routine amplification	TTATGCGGTATTA ATCTYCCTTT	*Clostridium* Group I	50 °C	~820	Franks *et al*., (1998)
Cbot 983		CARGRGATGTCAAGYCYAGGT				Van Dyke & McCarthy, (2002)
Cther 650	End point PCR/Routine amplification and qPCR assays	TCTTGAGTGYYGGAGAGGAAAGC	*Clostridium* Group III	60°C	~720	Van Dyke & McCarthy, (2002)
Cther 1352		GRCAGTATDCTGACCTRCC				
Clos 561	End point PCR/Routine amplification and qPCR assays	TAGAGTGCTCTTGCGTA	*Clostridium* Group IV	60°C	~580	Van Dyke & McCarthy, (2002)
Clept 1129		TAGAGTGCTCTTGCGTA				
Erec 482	End point PCR/Routine amplification and qPCR assays	GCTTCTTAGTCARGTACCG	*Clostridium* Group XIV	55 °C	~260	Franks *et al*., (1998)
CXIV 727		GTCCAGHARGYCGCCTT				This study
FibroQ153F	End point PCR/Routine amplification and qPCR assays	CCGKSCCAACGSSCGGHTAA	*Fibrobacter*	60°C	~104	McDonald *et al*., (2008)
FibroQ238R		CSCCWACTRGYTAATCRGAC				
FIB 1F	End point PCR/ production of template for qPCR	CCGKSCCAACGSSCGG	*Fibrobacter*	60°C	~855	McDonald *et al*., (2008)
FIB 2AR		ATCTCTCGCYGCGGCGWTYCC				

Primer Set	Purpose	Sequence 5'-3'	Target	Annealing temp.	Amplicon Size (bp)	Reference
1369F	qPCR assays	CGGTGAATACGTTCYCGG	general bacteria	60°C	~151	Suzuki *et al*., (2000)
Prok 1429R		GGWTACCTTGTTACGACTT				
NS1-Euk	End point PCR	CCAGTAGTCATATGCTTGTC	General Eukarya	50°C	~1600	White *et al*., (1990)
Univ 1390		GACGGGCGGTGTGTACAA				
T7--(dT)16	Priming production of first-strand cDNA from polyA+ RNA	TAATACGACTCACTATAGG$(TTTT)_4$	Non-specific for polyA+ RNA	N/A	N/A	Included with the MessageAmpII kit (Ambion)
T7-BpmI-(dT)16VN	Priming production of first-strand cDNA from polyA+ RNA	TAATACGACTCACTATAGGGAGA *GACCTC*$(TTTT)_4$VN	Non-specific for polyA+ RNA	N/A	N/A	Stewart *et al*., (2010)
Eub16s_27f	Production of 16S rRNA genes for subtractive hybridisation	AGAGTTTGATCCTTGCTCAG	General bacteria	55°C	~1300	Stewart *et al*., (2010)
Eub16s_1349rT7		GCCAGTGAATTGTAATACGACTCACTATAGGGACGGCTACCTTGTTACGACTT				
Eub23s_189f	Production of full-length 23S rRNA genes for subtractive hybridisation	GAASTGAAACATCTHAGTA	General bacteria	39°C	~2300	Stewart *et al*., (2010)
Eub23s_2490rT7		GCCAGTGAATTGTAATACGACTCACTATAGGGCGACATCGAGGTGCCAAAC				

Table 2.2 Details of primers used in this study

2.7 Double-stranded cDNA production for high throughput sequencing.

Two methods of dscDNA production were employed because the Just cDNA Double-stranded cDNA Synthesis kit (Agilent) was discontinued part way through the project.

2.7.1 Method 1: Just cDNA Double-stranded cDNA Synthesis kit.

First second strand cDNA synthesis reactions and final blunting of cDNA termini were carried out according to the instructions provided by the manufacturer. First strand synthesis was primed with 0.1µmol random hexamer primers and the starting material was 500ng of polyA+ RNA. cDNA was purified from the reaction mix by phenol:chloroform extraction and precipated with 2 volumes of ethanol and ¼ volume 3 M sodium acetate overnight at -20°C, or for 2 h at -80°C. Precipitated cDNA was collected by centrifugation at 17,000 x g at 4°C. The pellet was washed with ethanol, centrifuged again for 2 min and dried by vacuum centrifugation. dscDNA produced checked for integrity using agarose gel electrophoresis and quantified with the Qubit to accurately determine the quantity available for sequencing. After quality control, the material was stored at -20 °C until required.

2.7.2 Method 2: Superscript III and RevertAid Premium Double-Stranded cDNA Synthesis kit.

First strand synthesis was carried out using the Superscript III enzyme (Invitrogen). The reaction mix was made up according to the instructions provided by the manufacturer. The reaction was primed with 0.1µmol random hexamer primers and starting material was 500ng of polyA+ RNA. Second-strand synthesis was performed using the RevertAid Premiunm double-stranded cDNA kit (Thermo scientific), according to the instructions provided by the manufacturer. Material was stored and quantified as above.

2.8 Fosmid library preparation

Section 2.8.1 Buffer recipes for the HMW DNA extraction method.

The following buffers were prepared for use in the High molecular weight DNA extraction method for fosmid library production, based on the recipes from Neufeld *et al*. (2007).

SET buffer: 50 mM Tris-HCl, pH 8.0, 50 mM EDTA, and 0.75 M sucrose.

Fresh Proteinase K solution: 950 ml sterile water, 50 ml 1M Tris-HCl pH 8 and 20 mg proteinase K

Acidic pH 5.0 buffer: 35 ml 0.1 M-citric acid and 65 ml 0.1M-trisodium citrate.

CTAB buffer: A 2% (w/v) CTAB buffer was prepared from dilution of the CTAB stock solution as used in the Griffiths extraction method (10% (w/v) CTAB in 0.7 M NaCl solution and 240 mM potassium phosphate buffer (pH 8)) by dilution of 1 part stock solution to 4 parts sdd H_2O.

2.9 Bioinformatics and computational tools

Computational analysis of datasets generated from Illumina sequencing was carried out on a Mac Pro running Mac OSX version 10.6.8, with 8GB of RAM and four Intel Xeon processing cores each with a processing speed of 2.66 GHz.

BLAST searches were implemented with the command-line blastall program, version 2.2.26, against databases installed locally. Manipulation of sequence files in fasta and fastq formats was undertaken using the Bioperl toolkit (Stajich et al., 2002).

Chapter 3: A quantitative assessment of the composition of microbial communities in landfill leachate involved in cellulose degradation

3.1. Background.

Previous studies on the molecular ecology of cellulolytic microbes present in landfill leachate have used end-point PCR assays and the production of clone libraries to provide qualitative data on the indigenous taxa (Lockhart *et al*., 2006; Huang *et al*., 2005). It had been primarily established that various subgroups of clostridia (Collins *et al*., 1994) are predominant. PCR data using 16S rRNA genes as the target have also revealed that fibrobacters (Mcdonald *et al*., 2008) and anaerobic fungi of the *Neocallimastigales*, (Lockhart *et al*., 2006) can be detected, although not consistently between different sites. These findings suggest strongly that organisms which are close relatives of species well known as highly active cellulose degraders in the ruminant gut environment are also present in landfill. The mere presence of these organisms does not of course imply an important ecological function. Detection of the DNA of both fibrobacters and anaerobic fungi always required the use of nested PCR, which involves the amplification of (in this case) the 16S rRNA gene and subsequent second round amplification of rare 16S rRNA genes undetected by direct PCR. In the case of Neocallimastigales rRNA genes, detection was sporadic between samples and even this two stage nested PCR approach often failed to yield the specific amplification products. The overall picture appears to be that clostridia play a major role in cellulose breakdown in these environments and other rarer or less numerous organisms, somewhat related to better-known ruminant gut lineages, may form part of the community, but perhaps only a minor part, and their absence would not preclude the ability of the resident microbial community to degrade cellulose.

However, microbial ecological studies based merely on end-point PCR amplification and clone library production have limitations, are subject to bias, and, with the advent of new techniques, now somewhat obsolete. Quantitative PCR utilising SYBR green chemistry is capable of providing a much more sensitive detailed picture of the community structure, measuring not just presence/absence of a particular target gene but also the amount of amplification that occurs in a sample, and therefore the abundance of that gene.

By utilising a qPCR approach to study microbial communities it is possible to determine the proportion of the community rRNA genes that various taxonomic groups constitute. Here, total community DNA was used in qPCR assays where quantification was carried out with primer sets for 16S rRNA genes designed for either universal or taxon specific amplification. Results generated from these assays allowed the gene copy number of specific groups of bacteria to be compared to the total gene copy number of the 16S rRNA gene present in the total community DNA. This information gives an indication of the structure of the microbial community, given that organisms that are more abundant are significant contributors to the overall activity of the community and are major local occupiers of an ecological niche. This does not mean that groups detected at comparatively lower abundances do not have a role in the community; the recalcitrance of cellulose as a substrate favours synergy between multiple species, each employing their own array of enzymes. Numerically less abundant species might still be important in contributing to the overall community's ability to use cellulose as a carbon source. Amplification bias affects any attempt to investigate microbial ecology using qPCR. Efficiency of primer annealing is likely to vary between targets, with some sequences amplified more efficiently than others; it is important to take this into consideration when interpreting results from experiments of this type.

Here, a landfill environment was investigated in order to establish prevalence of know cellulolytic groups within the microbiota. Experiments were aimed specifically at enumerating bacteria from the *Clostridium* groups III, IV and XIV, and at the fibrobacters that may have been present.

3.2. Assessment of the suitability of primer sets for use in qPCR experiments

3.2.1 Control experiments.

It is important that qPCR primers are carefully designed as the technique requires a greater amount of stringency and precision than end point PCR. Group specific primers previously used for the detection of the taxonomic groups of interest here were not necessarily suitable for use in qPCR experiments. As a result, standard curves were produced to assess the performance of the primer pairs. Criteria for effective primer performance were (1) Generation of a single product of the expected size, (2) Ability to produce a reaction with an efficiency between 90 and 110%, (3) Production of a standard curve with an R^2 value of > 0.99, (4) have C_t values above the detection limit for the assay.

Single product production can be determined by generating a melt curve of the reaction product, and demonstrating the generation of a single peak. A melt curve with a large, single, sharply defined peak and a smaller more diffuse one at a lower temperature is also acceptable, the latter indicating dimerised primer sequence. Primer dimer is commonly visible at lower template dilutions and in negative control reactions.

The reaction efficiency % is calculated from the slope of the standard curve. An efficiency of 100% is ideal but difficult to achieve. Values between 90 and 110% are preferred. At 100% efficiency, the amount of product exactly doubles after each PCR cycle. The efficiency value therefore provides an indication of how well the primers are binding to the target; alternative binding sites may affect primer annealing efficiency such that one of the primer pairs is sequestered affecting efficiency of target amplification. A low efficiency could occur in the case where one primer binds to a non-target sequence, not causing the production of non-specific products on the one hand due to the specificity of the other primer sequence but still resulting in sequestering of one of the reaction components and a reduction in the efficiency of the actual target. Low efficiency can also indicate a degraded template, or the presence of chemical PCR inhibitors in the reaction mix. This last point is of particular concern when working with environmental DNA preparations.

If a primer set is reliable, it should produce a standard curve with data points that coincide closely for reactions where the same amount of template has been added. Although pipetting error both in production of the dilution series of the template and actual plate set up can introduce some variation, the coefficient of determination or R^2 value should be high, preferably above 0.99, as this indicates that the curve data fit very closely to a regression line.

The C_t is the threshold cycle where the increase in fluorescence as a result of amplification is detectable above the background, and is used to calculate the amount of starting template in each sample. C_t values can only be assumed as reliable if amplification for the reaction they are associated with occurs before the background amplification that will eventually be detected in the negative control wells. Reactions where amplification occurred at the same point or within three amplification cycles of the associated negative control were not considered reliable for the purposes of quantification.

Table 3.1 summarises the performance of the primers used here, as assessed by standard curve production. Some primers produced standard curves having lower R^2 and efficiency values than what was considered ideal, but better performance was obtained during later experiments. Only the original *Clostridium*

Group XIV set exhibited outstandingly poor results to an extent that required their replacement with a fresh set of primers tailored to use in qPCR experiments. Detailed primer characteristics are given in table 2.2.

Previously, primer sets had been developed and used for the detection of *Clostridium* groups by end point PCR (Van Dyke & McCarthy 2002), and for qPCR assays to detect *Fibrobacter* (McDonald *et al*., 2008). Details for all primer sets are listed in Table 2.2 It was expected that the primers designed for the detection of *Clostridium* groups by end point PCR might not perform well on a qPCR platform. All but one of the primer sets tested performed adequately, returning good melt curves with single product peaks and yielding standard curves of sufficient quality. The original primer set for the Group XIV Clostridia designed for end point PCR was not compatible with qPCR and produced a large amount of non-specific product. An alternative set of primers for this taxonomic group was therefore designed for use in qPCR experiments. This new set was assessed in the same manner as the previous (see Table 3.1) and found to perform adequately. The melt and standard curves for each primer set are presented in Figure 3.1.

Although the Neocallimastigales-specific primer set performed well, it was decided not to perform any further experiments on this target. This group was only rarely detected in environmental samples, and then only by nested PCR (McDonald *et al*., 2012).

3.2.2 Design of a new primer set for the 16S rRNA gene of *Clostridium* cluster XIV

qPCR primer design should ideally correspond to the following parameters according to the recommendations of Smith & Osborn (2009) and from product literature provided with Qiagen Quantifast reagents.

- 18-30 bp in length
- GC content 40-60%
- T_m for each primer should be as similar as possible
- Products smaller than 200bp
- Complementarity of 2 or more bases at 3' end avoided to minimise formation of primer dimer
- Mismatches of more than 3 bp between 3' end and target should be avoided
- GGG and CCC should be avoided at 3' end as these increase tolerance to mismatch

- No complementarity between regions of a primer and between different primers to avoid primer dimer and secondary structure formation.

In order to design a replacement set for the unusable Clostridium group XIV primer set, these primer sequences were used to retrieve matches from the target group of organisms from the Ribosomal Database Project (RDP) website using the probematch function. The Genbank accession numbers associated with these hits were then used to access and download Sequences from the Genbank database using the batch entrez facility at the NCBI website. These sequences were used to generate a multiple alignment using the Greengenes NAST server (DeSantis *et al*., 2006) and subsequently the program MUSCLE (Edgar *et al*., 2004) was used to tidy up inaccuracies in this initial alignment. This alignment formed the basis for primer design.

By visual inspection of the alignment, a region was found approximately 250 base pairs upstream of the binding site of the original forward primer. A new primer was designed to this region which met, or nearly met, the ideal properties of qPCR primers listed above. The difference in annealing temp was 2.2 °C (53.8 °C and 56 °C for the forward and reverse primer respectively) and the product size was approximately 260 bp.

The RDP database was used to ensure that the primers would not amplify targets beyond the taxonomic group they were designed to detect. It is likely that a given primer would match to 16S rRNA gene sequences outside of its intended target due to the high levels of conservation of this particular gene but specificity arises from the fact that unless both primers have corresponding non-specific targets, non-specific amplification will not occur. In this case, neither primer matched to the same 16S rRNA gene sequences of other taxonomic groups and therefore could be relied upon not to amplify DNA from species outside of the *Clostridium* XIV group.

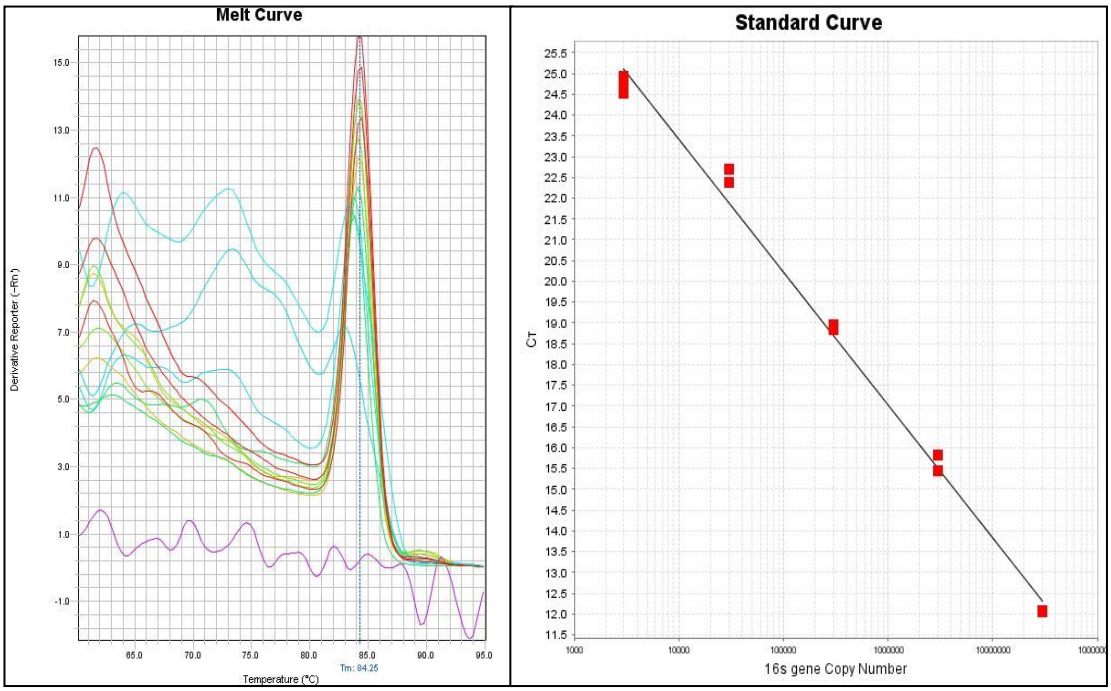
The convention for numbering of 16S rRNA gene primers corresponds to the position of the 5' end in the *E.coli* 16S rRNA gene. This value was determined for the primer designed here using the probebase primer searching facility (Loy *et al*., 2008) which returns this value in its search results. Hence, the naming of the primer “CXIV 727” denotes its taxonomic specificity and the position on the 16S rRNA gene that it targets.

Table 3.1: Summary of qPCR standard curve controls

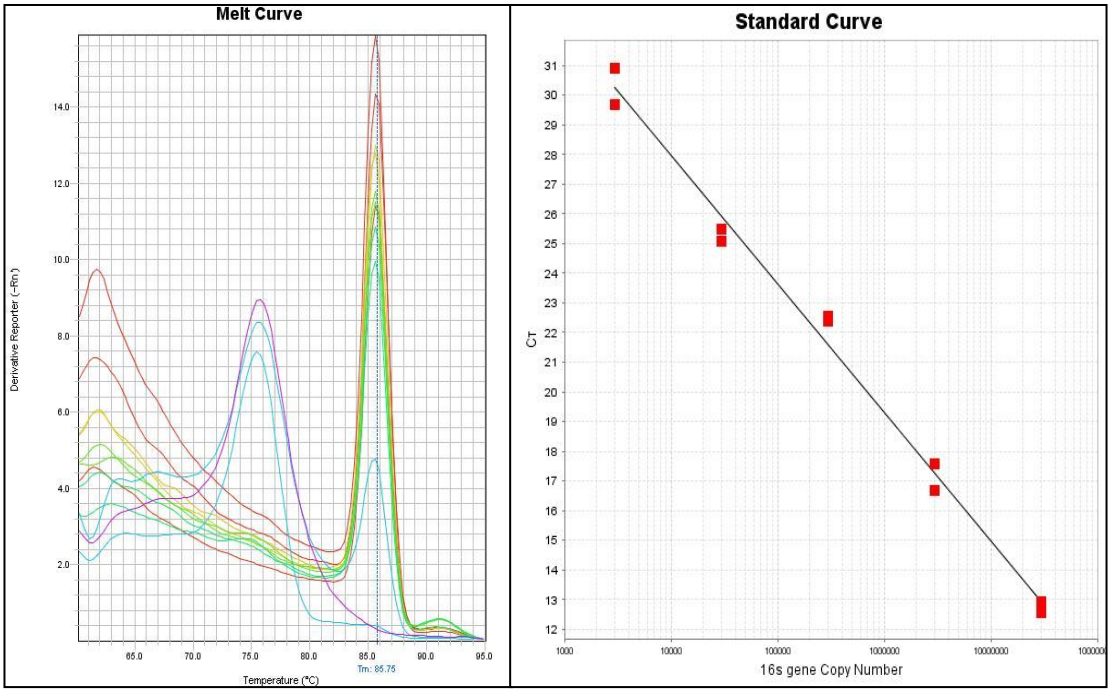
Primer set	Y - intercept	Slope	R^2 Value	Efficiency (%)

target				
General Bacteria	36.248	-3.203	0.994	105.237
Clostridium group III	45.285	-4.327	0.991	70.258
Clostridium group IV	36.75	-3.539	0.989	91.668
Clostridium group XIV (original set)	37.175	-0.981	0.128	946.106
Clostridium group XIV (redesigned)	38.548	-3.361	0.998	98.395
Fibrobacter	52.504	-4.104	0.978	75.241

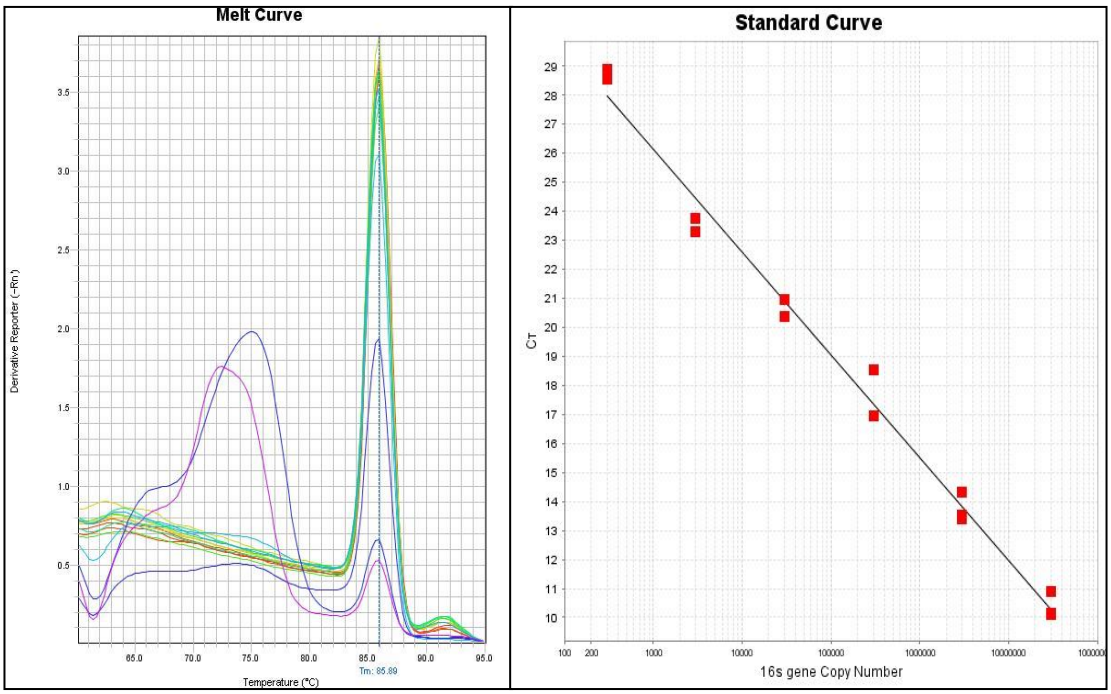
A. General Bacteria primer designations



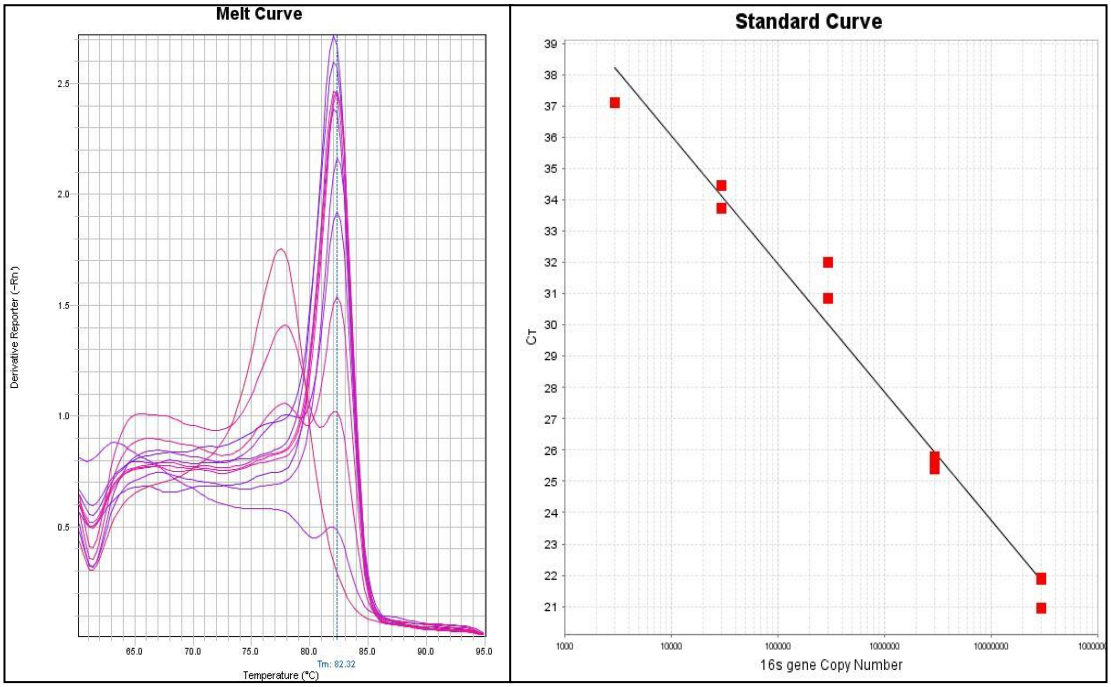
B. Group III clostridia primer designations



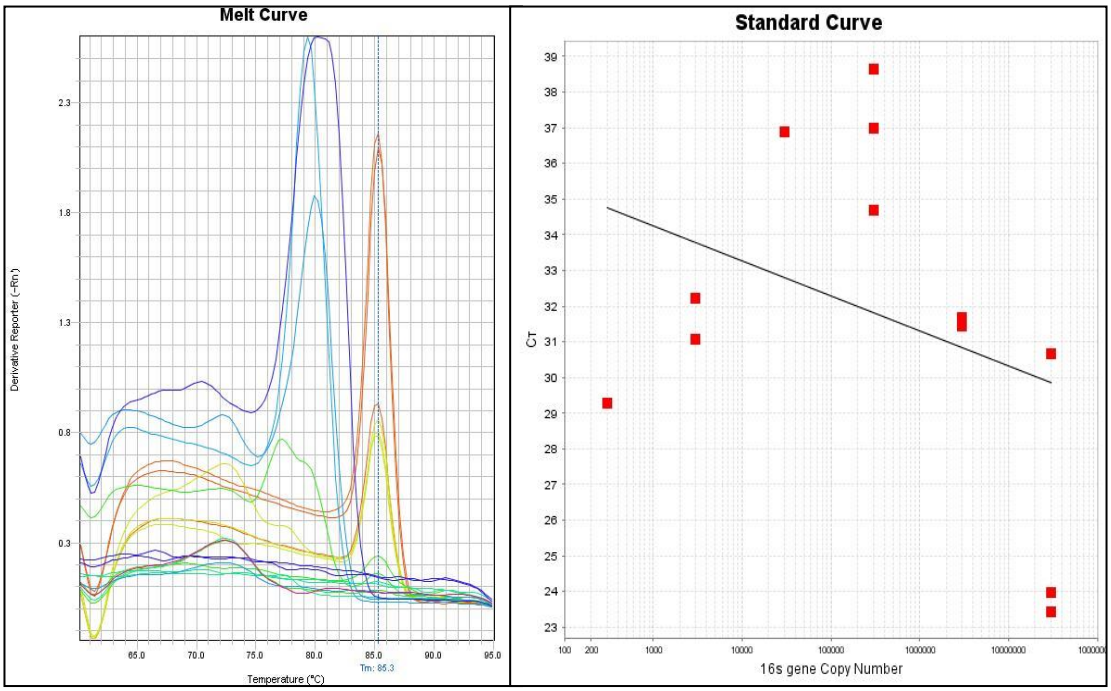
C. Group IV clostridia primer designations



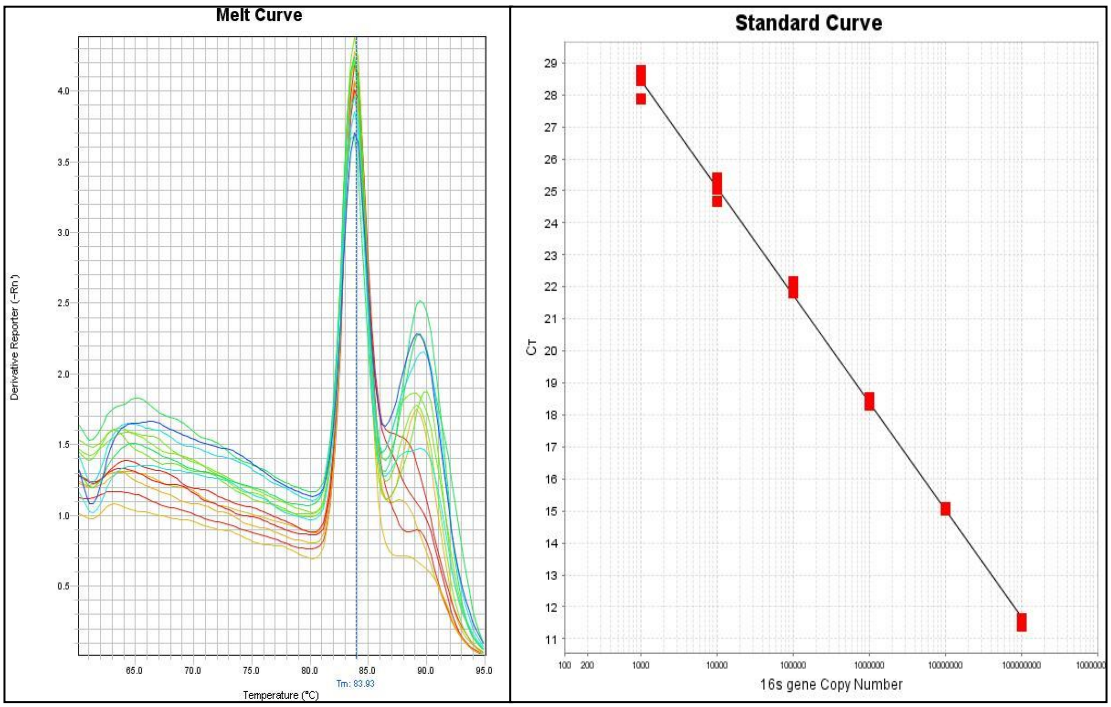
D. *Fibrobacter spp.* primer designations



E. Group XIV clostridia – original primer set primer designations



F. Group XIV clostridia – redesigned primer set primer designations



G. Neocallimastigales (anaerobic fungi) primer designations.

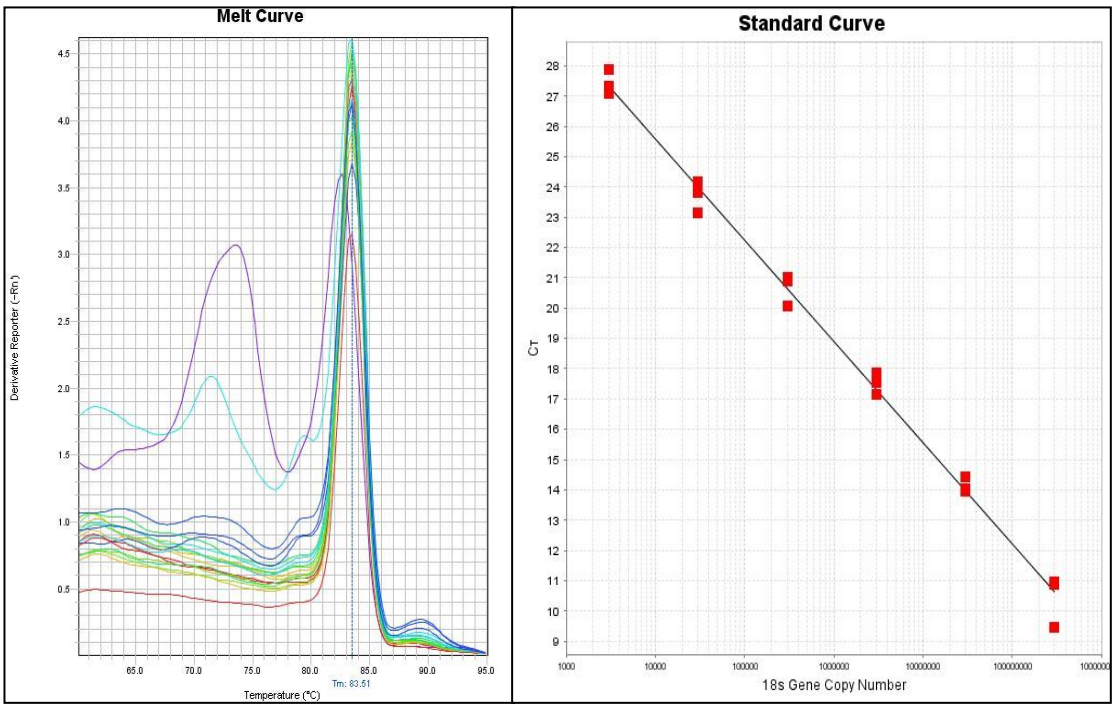


Fig 3.1: Melt curves and Standard curve plots generated from control assays to test the performance of primer sets prior to quantification of unknown samples.

Numerical values for the standard curve plots are summarised in Table 3.1. Melt curves were generally characterised by one large, single spike peak, with smaller more diffuse peaks appearing at lower temperatures, associated with lower template dilutions and often coinciding with background detection from the negative control assays, . Generally the performance was considered good or satisfactory for all primer sets, the data points plotted all distributing near or close to the regression line, with the notable exception of the original group XIV *Clostridium* set. In this case, performance was very poor but the design of a replacement set resulting in an adequately functioning primer pair.

3.3 Quantitative assessment of cellulose degrading bacteria from landfill leachate samples and colonised cellulose: Experimental design and setup.

Samples of total community DNA for this study were either previously available from a study by McDonald *et al* (2008), or derived from colonised cotton baits incubated in microcosms containing landfill leachate from the Bromborough Dock landfill site (C.f. section 2.1.2). Two microcosms were used, microcosm one containing leachate from Bromborough Dock risers 3 and 4, and microcosm 2 containing leachate from riser 5. Cotton baits for the qPCR study were incubated in microcosms for 2 weeks. In addition to community DNA, RNA extracts were also obtained from the colonised cotton samples. The RNA was reversed transcribed and the resulting cDNA was used as template material in qPCR experiments. The DNA samples from Bromborough dock risers 1-5 inclusive and DNA and RNA extracted from colonised cotton from the two risers produced 9 samples in total.

The primers found to be suitable for use in qPCR assays (Table 3.1) were used to carry out quantitative assessments of the microbial communities of landfill leachate samples and communities associated with biofilms on cotton baits incubated in leachate.

Each group specific primer set was used in a separate experiment, on a separate 96 well plate, where standard curves for both the specific and universal bacterial primers were generated simultaneously in triplicate from the same dilution series of control template DNA; each unknown sample was assayed with both the general and specific set, again in triplicate. This design allows for ratios between gene copy numbers generated with the general and specific primer sets to be produced while minimising the effect of experimental error. A schematic detailing the plate set up of the qPCR assays carried out here is presented in Table 3.2. Standard curves generated from the experiments and used for the production of quantitative abundance data are presented in Fig. 3.2.

Assays deemed acceptable for generation of quantitative data had R^2 values >0.99 with one exception for the CXIV assay where the general bacterial standard curve had an R^2 value of 0.972, but as the specific primer set curve had an R^2 value of 0.998 it was decided not to repeat the assay This value demonstrated good performance of the specific set, and the general bacterial primers had performed well elsewhere. Efficiencies were between 82.2% and 104.9%.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Standard Curve for quantification – General Bacteria primer set. Starting from 3 x 10^8 gene copy number ul^{-1} and descending with a 1:10 dilution factor to 3 x 10^3; curves performed in triplicate.		3 x 10^8	Standard Curve for quantification – Specific primer set layout as for general bacteria curve.		3 x 10^8	Bromborough Dock Riser 1 DNA: Triplicate samples			Bromborough Dock Riser 1 DNA: Triplicate samples		
B			3 x 10^7			3 x 10^7	Bromborough Dock Riser 2 DNA: Triplicate samples			Bromborough Dock Riser 2 DNA: Triplicate samples		
C			3 x 10^6			3 x 10^6	Bromborough Dock Riser 3 DNA: Triplicate samples			Bromborough Dock Riser 3 DNA: Triplicate samples		
D			3 x 10^5			3 x 10^5	Bromborough Dock Riser 4 DNA: Triplicate samples			Bromborough Dock Riser 4 DNA: Triplicate samples		
E			3 x 10^4			3 x 10^4	Bromborough Dock Riser 5 DNA: Triplicate samples			Bromborough Dock Riser 5 DNA: Triplicate samples		
F			3 x 10^3			3 x 10^3	Microcosm 1 DNA: Triplicate samples			Microcosm 1 DNA: Triplicate samples		
G	Negative control: 3 wells			Negative Control: 3 wells			Microcosm 1 cDNA: Triplicate samples			Microcosm 1 cDNA: Triplicate samples		
H	Microcosm 2 DNA: Triplicate samples			Microcosm 2 DNA: Triplicate samples			Microcosm 2 cDNA: Triplicate samples			Microcosm 2 cDNA: Triplicate samples		

Table 3.2: Layout of qPCR plates used in the quantification experiments.

The experimental design also allowed spurious results to be removed, while maintaining a quantification based on an average of two remaining assays. This allowed for the removal of apparent outliers from the data in cases where one Ct value out of a set was wildly different and displayed an abnormal melt curve, amplification profile or was otherwise flagged by the Steponeplus software (Applied Biosystems).

3.4 Relative abundance of cellulolytic taxa in landfill leachate and colonised cotton samples.

Relative abundances generated for *Clostridium* Groups III, IV and XIV, and *Fibrobacter spp.* demonstrate the occurrence and distribution patterns of these organisms. The *Clostridium* groups were quantified on the basis of good quality standard curves and reproducible results, with only occasional outliers that required removal.

The data for the *Fibrobacter* assays was challenging to analyse. The first plate yielded standard curves for both bacterial and *Fibrobacter*-specific primers which were substandard. Additionally, the majority of the reactions had Ct values below the sensitivity of the assay, with amplification occurring concurrently with the no template control reactions. Only two samples appeared to report genuine signals, DNA extracted from Bromborough Riser 4 leachate and DNA from microcosm 1 which contained leachate from Riser 4 in any case. This did agree with previous attempts to investigate presence/absence of fibrobacters using end point PCR as they were only previously discovered by nested amplification in leachate from riser 4. McDonald *et al*. (2012) report that Fibrobacters were detected through nested PCR from landfill leachate samples routinely but that only DNA Bromborough Dock Riser 4 ever yielded a direct product so it is possible that the qPCR assay would only be able to detect fibrobacters from these samples. Repetition of the experiment produced better quality standard curves and it was once again observed that Bromborough Riser 4 and microcosm 1 produced genuine signals, whereas all other samples exhibited amplification at the same, or within three cycles of, the negative control. The reproducibility of this observation indicates that the significant detection of *Fibrobacter* DNA in only two samples was not due to experimental error but represented an authentic result.

By the time the second qPCR experiment was carried out, the template material for the unknown samples had been through multiple freeze-thaw cycles and may have begun to deteriorate as a consequence, which may have contributed to

the large standard deviations in the triplicate repeats. It was decided to determine *Fibrobacter* abundance using Ct values from the first plate, but for the production of data sound enough for publication (McDonald *et al*., 2012), quantification was achieved with a standard curve from a previously published study (McDonald *et al*., 2008). The standard curves are presented in Figs. 3.2 and 3.3.

The abundance, as a percentage of the total abundance of total bacteria, of the *Clostridium* groups in each sample is presented in Figs. 3.4 – 3.6. The abundance of the *Fibrobacter* is displayed in Fig. 3.7, which is a simpler comparison between the only two samples in which the organism could be reliably detected.

3.4.1 Group III clostridia

For the group III clostridia, the relative abundance ranged from 1.05 to 7.84% across the leachate samples (Fig 3.4). This indicates that the group is a common resident of landfill leachate, and is occasionally numerous.

In microcosm 1, where a highly active biofilm rapidly degraded the cotton bait, the group appeared to be an important component of the community comprising 17.4% of the total population by analysis of the DNA (Fig 3.5). The group III clostridia appeared to be enriched in the biofilm associated with the cotton when compared to the quantities of these organisms from leachate samples taken from the original source of the microcosm 1 leachate. The relative abundance detected by qPCR on an RNA sample reverse-transcribed to cDNA from the microcosm 1 put the abundance at 53% which could be an indication that the group III clostridia are highly metabolically active in this environment, but the data based on RNA samples may be inaccurate and should not be used to draw conclusions. In addition to the assay for this sample yielding an abnormal melt curve, determination of an abundance of 0.14% in microcosm 2 for DNA but 36.29% for RNA strongly suggests that the RNA values are somehow artificially inflated. Values produced from RNA samples were not used for publication or to draw any of the conclusions discussed in this chapter.

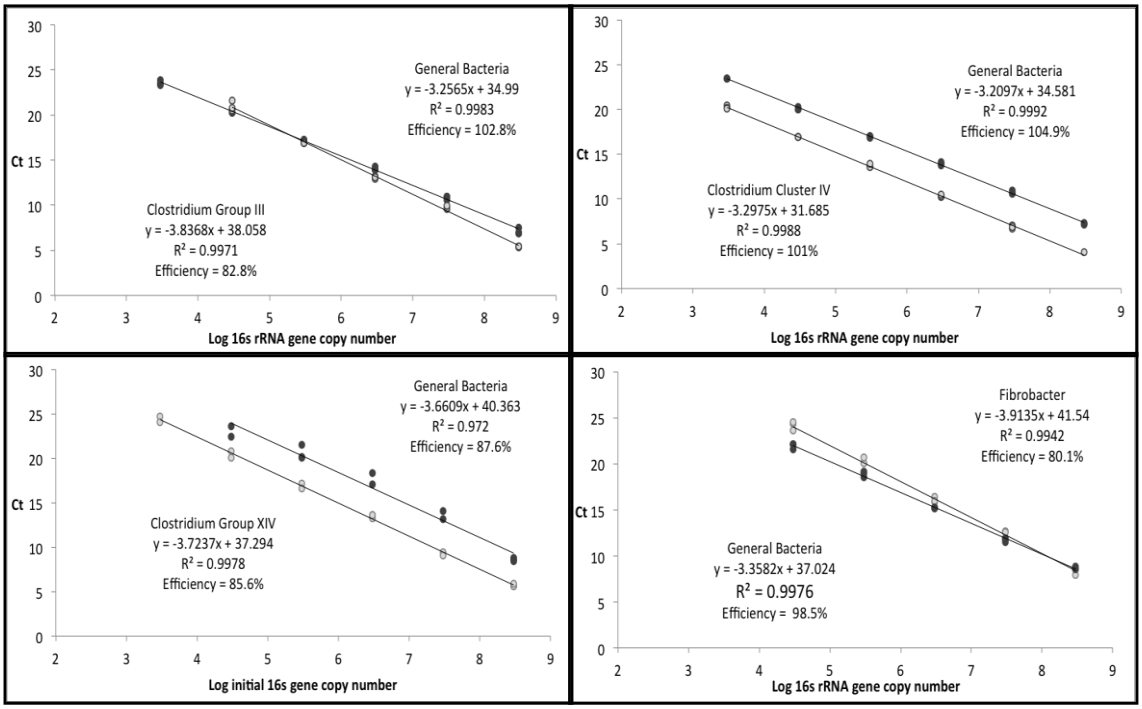
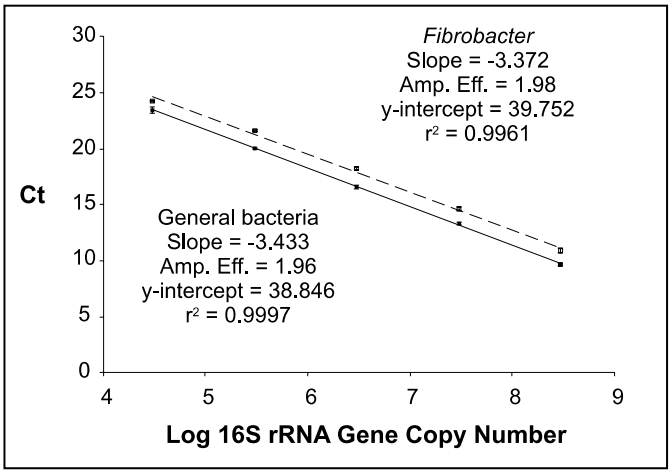


Fig 3.2 Standard curves produced for quantification experiments

Figure 3.3 Standard curve for Fibrobacter, reproduced from McDonald *et al*. (2008).



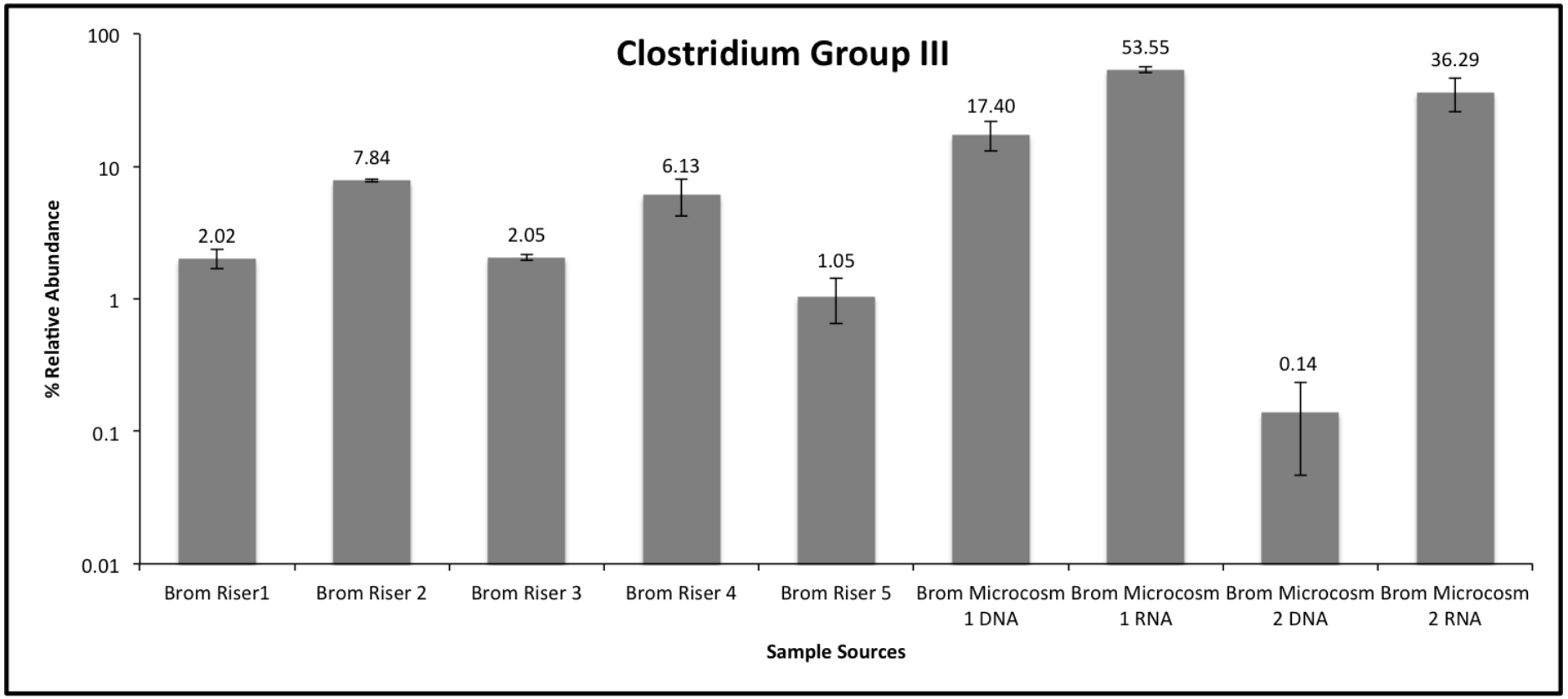


Figure 3.4 Relative abundance of the *Clostridium* group III across all the samples investigated in this study. Error bars represent standard error of the mean for either three or two relative abundances calculated from separate assays parallel assays for each unknown sample, Representing a measure of reproducibility and reliability for the qPCR assays.

3.4.2 Group IV clostridia

The group IV clostridia abundance for leachate samples and microcosm DNA samples ranged from 0.13 to 3.21%. Even in the samples where this group was found at the highest levels, the group IV clostridia do not seem to be very important members of these communities. Their numbers were elevated in the biofilm community from the colonised cotton in microcosm 1 with respect to the leachate samples but here they still comprised only 3.2% of the total population.

3.4.3 Group XIV clostridia

For the group XIV Clostridia a reasonably high abundance stands out in the case of one of the leachate samples from riser 3, where they are apparently a numerous part of the total community, comprising 12.99% of the total population. For all other leachate samples and microcosm DNA samples the abundances range from 0.51% to 2.89%, which demonstrates that this group is not usually a major component in these anaerobic environments, but in riser 3 whatever local conditions prevail are apparently more favourable.

3.4.4 *Fibrobacter*

The fibrobacters were detected in only one of the leachate samples, specifically from riser 4, where they are a small part of the total community, but appear to be enriched in microcosm 1, where they were calculated to comprise as much as 28.8% of the total by one method employed (the standard curve used in McDonald *et al*. 2008) and nearly 7% using the curve produced in this study. This suggests an overall rare presence is the norm for fibrobacters in landfill microbiota but they might be found to be locally numerous as colonisers of cellulosic material within the site.

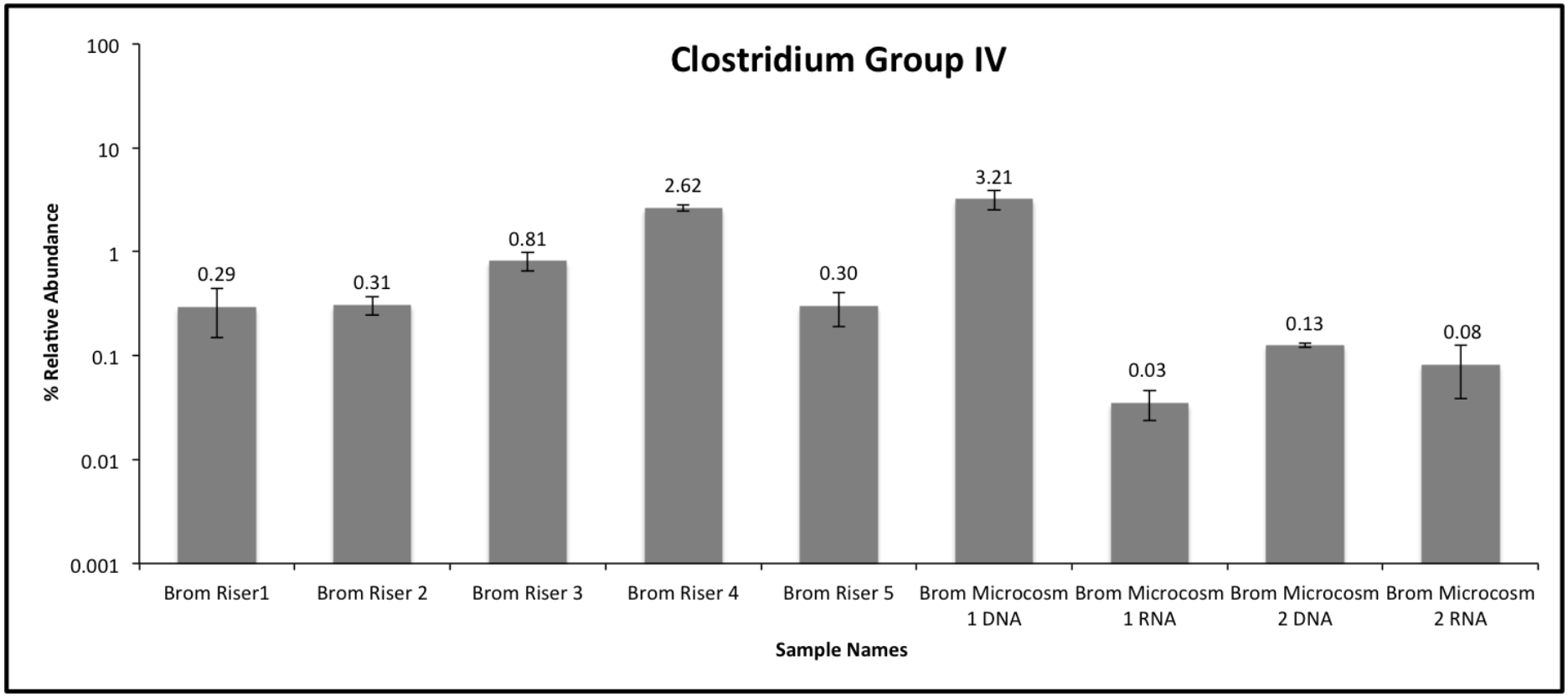


Figure 3.5 Relative abundance of the *Clostridium* group IV across all the samples investigated in this study. Error bars represent standard error of the mean for either three or two relative abundances calculated from separate assays parallel assays for each unknown sample, representing a measure of reproducibility and reliability for the qPCR assays.

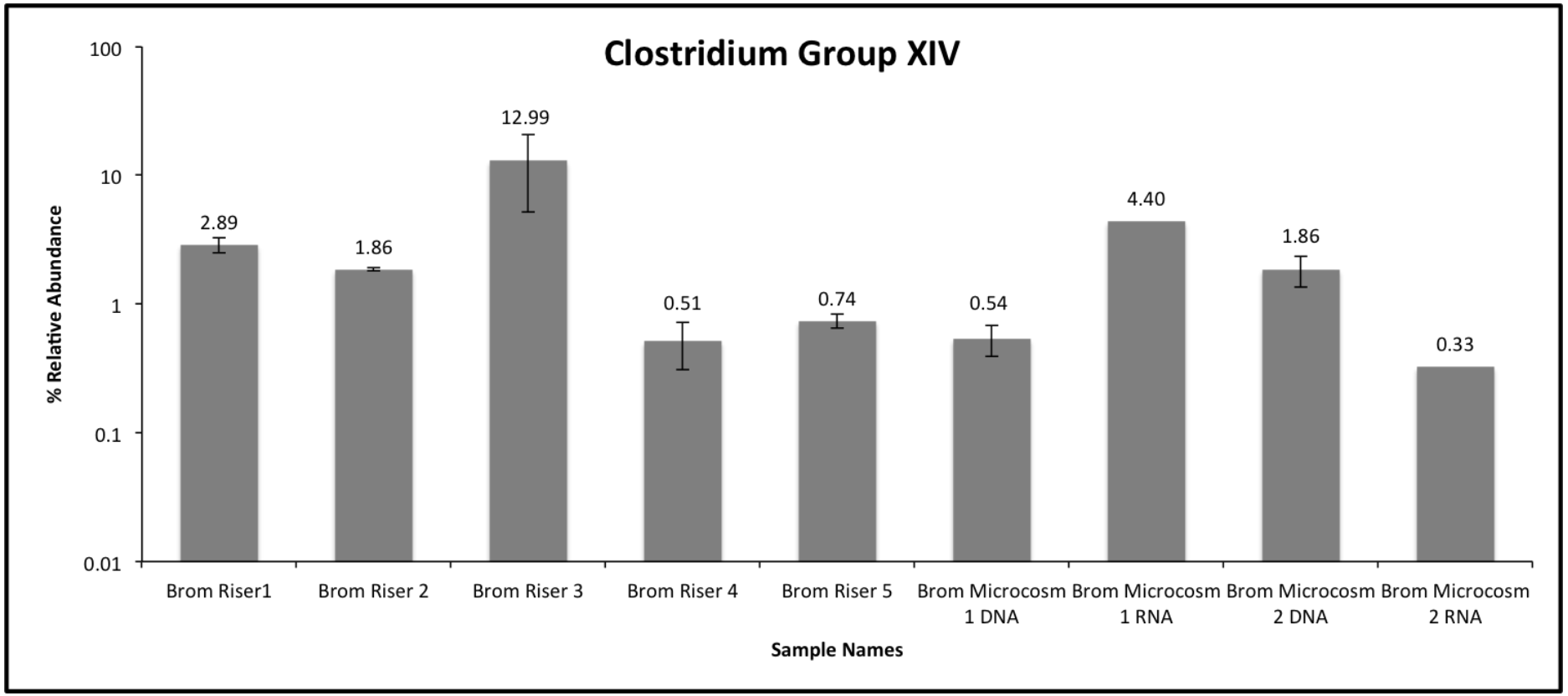


Figure 3.6: Relative abundance of the *Clostridium* group XIV across all the samples investigated in this study. Error bars represent standard error of the mean for either three or two separate relative abundances. In two cases (Brom Microcosm 1 RNA, Brom Microcosm 2 RNA) error bars are omitted, as only one sample of the triplicate repeats constituted a reliable data point and therefore standard deviation could not be calculated.

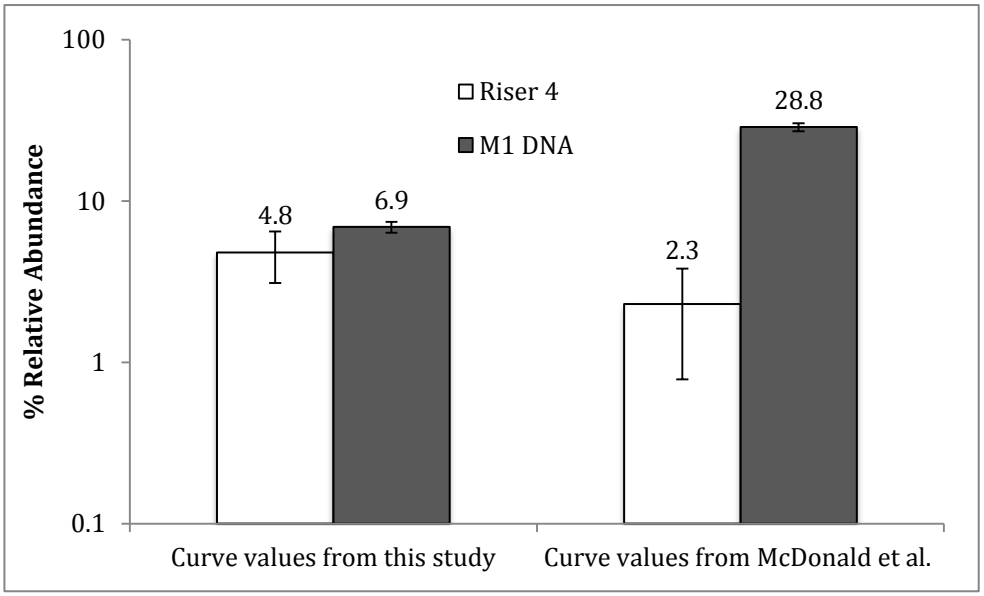


Figure 3.6: Relative abundance of *Fibrobacter* from both samples where a reliable signal could be obtained, and quantified separately from two different standard curves (presented in Figs 3.2 and 3.3). Error bars represent error of the mean for either three or two relative abundances calculated from separate assays parallel assays for each unknown sample, representing a measure of reproducibility and reliability for the qPCR assays.

3.4.5 Localisation of cellulose-active lineages in landfill to cellulosic material.

The availability of DNA extracted directly from leachate from Bromborough Dock landfill site, and the availability of landfill leachate microcosms enabled comparison of *Fibrobacter* abundance in the general bacterial population and in a biofilm population colonising crystalline cellulose, in the form of cotton, incubated in leachate. In this case, the two microcosms also happened to produce an interesting comparison between two very different colonising biofilms, one highly active against the crystalline cellulose source it had colonised and the other with no degradative activity at all. This comparison yielded the most interesting insights which can be gleaned from the results of this study. Fig 3.7 below summarises the results of this analysis.

Microcosm 1 contained leachate from risers 3 and 4, and riser 4 was the one area in the landfill site where *Fibrobacter* was readily detected. *Clostridium* group XIV was very numerous in riser 3 but was not found to be such an important coloniser of cotton in microcosm. The fibrobacters were found to be highly enriched in the colonised cotton however and it seems likely that the presence of these, along with the large number of representatives from the group III clostridia, was responsible for the high level of cellulose degradation observed in this microcosm (Fig 3.7).

Microcosm 2 contained leachate from riser 5. No fibrobacters were detected here and levels of all *Clostridium* groups were low. The cotton in this microcosm exhibited no apparent degradation and it appeared that an ability of the colonising microflora to breakdown cellulose was utterly lacking.

Comparing these two results it does appear that the presence of cellulose-degrading groups, known for their role in the rumen, does determine whether or not environmental communities will or will not have a cellulolytic ability. It also appears that groups of known cellulolytic organisms from the rumen have environmental relatives. This was previously known to be the case for the clostridia which are rather ubiquitous in nature but is a novel finding regarding the fibrobacters, thought to be specialised rumen symbionts previously. Their detection in this study demonstrates that they can be present in high numbers in microbial communities that colonise cellulosic material and exhibit high levels of degradative activity.

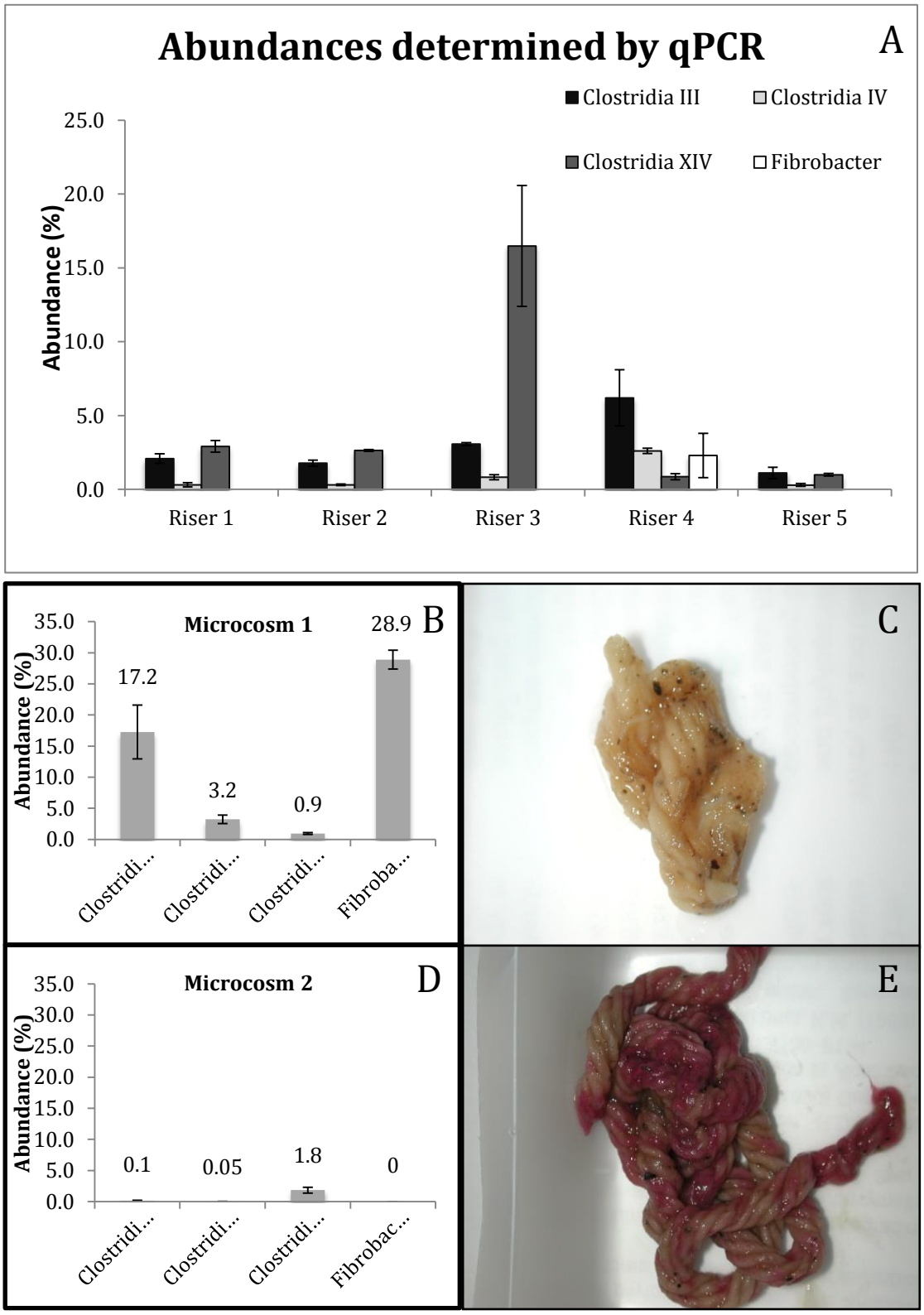


Figure 3.7 qPCR analysis of clostridia and *Fibrobacter spp*.in landfill leachate and colonised cotton; full legend continued on next page.

Figure 3.7 (cont.) (A) Abundance of clostridia and *Fibrobacter spp*. across five leachate samples determined by qPCR. (B) Abundance clostridia and *Fibrobacter spp*. detected by qPCR in the biofilm colonising cellulose biats in microcosm one which contained leachate from risers 3 and 4; the cotton bait itself is illustrated in (C). (D) Shows the abundances in microcosom 2, containing leachate from riser 5, and (E) shows the cotton bait from this microcosm, which appears from visual inspection and qPCR data to be colonised by a markedly different community to that in microcosm 1.

3.5 Discussion

The work in this chapter was undertaken to shed light on the ecology of the cellulose-degrading bacteria, focussing specifically on *Clostridium* groups III, IV and XIV and the genus *Fibrobacter*. A presence of these groups had been pre-established in the leachate of Bromborough Dock landfill site (McDonald *et al*. 2008). Using qPCR, their contribution to the overall community was further assessed.

The performance of all primer pairs used in the study was first tested in control reactions and in the first instance some low efficiencies were encountered. These initial borderline performances were thought to be partly due to old reagents, which were replaced before performing the final experiments for quantification. As a result general improvements in efficiencies and R^2 values were obtained.

Primer design to achieve specific amplification of the 16S rRNA gene of taxonomic groups is difficult with the 16S rRNA coding region consisting of very highly conserved regions, interspaced with variable sections. There is therefore a challenge involved in designing primers specific enough to only amplify a certain group comprehensively, taking in all of the diversity which may be present within it and simultaneously excluding all other bacteria. Primer design carried out here aimed to avoid non-specific amplification by producing a primer pair where although each individual primer might have some background non-specific binding capacity there was no non-specific binding to genes outside of the target group common to both member of the pair. In this way, specificity was achieved for the pair as a whole whilst tolerating small amounts of non-specificity in the design of individual primers, which allowed guidelines or qPCR primer sets to be adhered to.

The relative abundance calculations used here to establish the proportion of the target groups were intended to circumvent some of the controversy over the use of absolute numbers in qPCR experiments. The use of this relative quantification method does however rely on an assumption that is not entirely valid, which is that

the numbers of rRNA operons in all species in the community are the same. Two organisms equal in terms of the number of cells present might appear to be differently abundant if one possesses two extra rRNA gene copies. Variation in rRNA gene copy number probably does have an influence on the data obtained to some extent. With sensible interpretation of the results, this should not represent a fatal flaw in the experiment. Common sense dictates that only major differences in the abundance of two different groups in the same sample represents a remarkable result and comparisons between the abundances of the same group over different samples is unaffected.

It is also inaccurate to regard the abundances of the target groups determined as true abundances of the total bacterial population as there is no such thing as a truly universal primer set and the general bacteria primer set used here will really only target a subset of the total population. However, as all abundances are calculated in the the same way, all abundances are based on the same subset of the total bacteria. This is therefore a consistent artefact and does not really affect the data.

It seems clear that the biofilm which colonised the cotton in microcosm 1 is significantly better at breaking it down and that the community has a much higher proportion of cellulotrophic organisms. Why it should be that one sample of cotton was colonised and broken down rapidly and why another was colonised by a community effectively unable to do so is less clear. It seems to be the case that colonisation by the key genera was blocked somehow. This could be directly related to the composition of the community. It is possible that cellulolytic microorganisms are not present in this particular leachate sample, or form just a small part of community and this precluded them from gaining a foothold on the cotton which was rapidly colonised by non-cellulolytic species. The presence of non-cellulotrophs may well have rendered the cellulosic material inaccessible to species actually capable of breaking it down and using it as an energy source. Landfill sites are fundamentally heterogeneous and it is plausible that sampling different sites (or different, physically separated, areas of the same site in this case) could reveal very different components of the microbial population, as was observed here. It can be hypothesised that conditions in the area sampled by riser 3 were more favourable to the growth of cellulotrophs, perhaps due to a greater quantity of cellulosic waste present in this area, perhaps for other reasons. The observations do paint a very neat picture; a crystalline cellulose source, colonised by specific genera know to be highly active cellulose degraders was rapidly digested and an identical crystalline

cellulose source which was not colonised by these genera was not at all degraded over the same time span.

Fibrobacter quantities were effectively undetectable for most of the samples which correlates with previous observations, increasing credence that these results are painting an accurate picture. The overall view proposed is that although fibrobacters may not be predominant, their presence does imply that a high level of degradative activity against cellulose and cellulosic materials will be exhibited by that community. *Clostridium* group III and fibrobacters were both enriched in the highly active biofilm of microcosm 1 but *Fibrobacter* numbers were the highest, which could mean this group is the most important cellulose degrader.

The illumination of an important role in environmental cellulose degradation played by a genus previously though to be an exclusive inhabitant of the digestive tract of ruminants is the key finding in this qPCR survey. This observation suggests certain other implications. First, given that the genus *Fibrobacter* is known to exhibit potentially novel cellulases (Wilson, 2009), there may be biologically and commercially interesting enzymes produced in other anaerobic environments. Second, there may be other cellulose degrading species present in this environment, and others, carrying out cellulose degradation that remain unknown.

Until recently, *Clostridium* lineages were thought to be mainly responsible for anaerobic breakdown of cellulose outside of the gut, where *Fibrobacter* and anaerobic fungi are also heavily involved in the process. This was a logical deduction given that clostridia have been found ubiquitously in such environments and fibrobacters have not (Ransom-Jones *et al*. 2012). *Fibrobacter* DNA had been demonstrated to be difficult to amplify (Tajima *et al*. 2001); detection of this genus by molecular means is therefore difficult and so with respect to the more widespread distribution of *Fibrobacter* across anaerobic environments outside of the herbivore gut, an absence of evidence is not evidence of absence. The work carried out here has established that *Fibrobacter spp*. can be present in very high numbers in biofilms colonising cellulosic material in anaerobic landfill leachate and these organisms could to be important contributors to cellulose breakdown in landfill.

Given that *Fibrobacter* DNA is difficult to amplify it may be that isolating community RNA and reverse transcribing the 16S rRNA gene to produce a cDNA template will yield a clearer picture of its presence. This approach was used here when investigating the microcosms, but the cDNA reactions in the qPCR assays produced quantifications with very high standard deviations, and in many cases only one out of the three triplicate repeats produced a melt curve that indicated a reliable reaction had taken place. On the whole, the data based on cDNA amplifications was

not considered reliable and no true detection of *Fibrobacter* was recorded by this method. It is possible that this problem was caused by the age of the samples and that the RNA extracts obtained had suffered degradation. This may have influenced downstream qPCR amplification and performance. *Fibrobacter* was detected in high levels based on DNA amplification from microcosm 1 in any case so it is possible that if much higher numbers of the organisms are present, poor DNA amplification efficiency is less relevant.

This work used qPCR to demonstrate a previously unappreciated ecological role for the *Fibrobacter* genus outside of the rumen. It also demonstrated the efficacy of cellulose cotton baits as a method for enriching for cellulose degrading organsisms in the environment. Based on these findings the use of cellulose baits and a less targetted approach to community analysis (e.g. metagenomics and metatranscriptomics) in order to gain a more complete picture of the microbiota present in the sample and investigate the functional genes behind the breakdown of cellulose are a promising line of research for future work.

Chapter 4: Production of a metatranscriptome and a fosmid library from environmental DNA.

4.1 Background

In order to investigate the microbial community of a lake sediment, two separate strands were developed. One aspect of the work was based on high-throughput sequencing, with the goal of obtaining an enriched sample of community mRNA which could be used to produce DNA libraries for 454 pyrosequencing. The relatively long reads produced by this sequencing platform, when compared to other high-throughput technologies, would make for easier classification and identification of any reads corresponding to known proteins, especially glycosyl hydrolases and other proteins implicated in cellulose degradation. The resulting metatranscriptome dataset would therefore serve as a source of information on the active members of the microbial community in the lake sediment and the families of glycoside hydrolase genes being expressed. Production of a pyrosequenced metagenome in parallel is by comparison relatively straightforward.

Parallel to the sequencing effort, cloning of high molecular weight DNA was to be carried out in order to generate a fosmid library. Fosmids containing inserts of 30-40kb in length are likely to contain entire genes or even sets of genes and the cloning of metagenomic DNA from colonised cotton embedded in the lake sediment was planned to generate a library that could be screened for cellulases by expression and PCR/oligonucleotide probing. Basic expression screening using an agar plate based approach has yielded positive results elsewhere (Liu *et al*. 2011, Geng *et al*. 2012). Hit rates for discovery of glycoside hydrolase enzyme families and related organisms in environmental metagenomes have, however, generally been very low (Kakirde *et al*. 2010). The use of cotton cellulose baits *in situ* has been shown to exert an enrichment effect and increase the yield of cellulase fragments in a pyrosequenced metagenome (Edwards *et al*. 2010). Information from the metatranscriptomic sequencing could be used to design PCR primers or oligonucleotide probes to specific genes discovered by bioinformatic analysis of the dataset. These molecular tools could be employed to screen the fosmid library further with advantages over expression screening methods. Genes may require, in addition to their simple transcription and translation, specific secretion pathways or else an assemblage of other proteins (e.g. cellulosomes) in order to carry out their function. Moving beyond culture-based screening into molecular methods allows

otherwise undetectable genes to be discovered within the fosmid library. Expression screening would not be able to detect enzymes dependant on secretion processes absent in an *E. coli* fosmid host strain or proteins with non-catalytic functions such as carbohydrate binding domains. Clones harbouring genes of interest can be used to produce large quantities of protein for biochemical characterisation and functional analysis.

In order to generate a metatranscriptome and fosmid library, it was necessary to employ a variety of techniques and protocols to ensure the generation of high-quality data from these samples. For example, extraction of nucleic acid from cellulose cotton baits in good yields is difficult and in the past the method of Griffiths *et al*. (2000) has been a solution to this problem, and was perfectly adequate for the extraction of RNA from the colonised cotton baits. The relatively violent processing (bead beating) of the sample did cause a great deal of shearing of DNA preparations however and it could not be used at all for the production of high molecular weight fragments of metagenomic DNA for which other extraction methods were required. This necessitated methods comparison and optimisation to generate sufficient high molecular weight DNA of the required purity.

4.2 Extraction of RNA from cotton cellulose baits for metatranscriptomic sequencing.

4.2.1 Community RNA extraction

The method of Griffiths *et al*. (2000) was effective when employed as an extraction method for producing high quality RNA preparations of an acceptable purity and quantity for use in subsequent experiments. There was variation between samples but this is to be expected when working with heterogeneous environmental material. Multiple preparations were made from cotton bait samples from each sampling point, as described in section 2.1. Extracts exhibiting high levels of contamination (i.e. low or excessively high $A_{260/280}$ and $A_{260/230}$ ratios) were discarded. Two extracts from each sampling point were selected to be combined, forming a stock of RNA to be used for metatranscriptomic sequencing. The quality of the RNA aliquots was checked before pooling; their profiles are summarised in Table 4.1.

Table 4.1 Yield and purity of RNA obtained from colonised cotton baits using the extraction method of Griffiths *et al*. (2000). Extractions were pooled into a single aliquot which was used as source material for further experiments.

Week of sample	**Nanodrop concentration (ng ml^{-1}).**	**Qubit concentration (ng ml^{-1}).**	**A260/280**	**A260/230**
Week 2	18	19.2	2.16	1.45
Week 2	16	18	1.97	1.96
Week 4	18	21.2	2.37	1.89
Week 4	71	23.1	2.13	2.56
Week 8	111	66.2	2.07	2.58
Week 8	156	67	2.09	2.42

Concentrations reported here were consistently lower from the Qubit platform. This is normal as the nanodrop is a great deal more sensitive to background contamination in the sample and to trace amounts of DNA remaining after the DNAase treatment. The RNA preparations analysed here were put through one more cleanup step before being further processed. Yields were increased for the extractions from cotton which had been incubated in the lake for longer (Table 4.1.) which is not surprising since there was more time for colonisation to occur on the surface and therefore more cells present and more nucleic acid available for harvesting.

The RNA obtained was deemed suitable for use with the MessageAmp kit, in order to produce amplified RNA which in turn could be used to generate sufficient cDNA for the production of a library for 454 sequencing. The pooled RNA stock was kept at -80^{o}C until use and stored overnight at -20 °C during intermediate steps in the protocol. RNA extracts from all three sampling points were stored frozen for further use at a later date if needed.

4.2.2 Preparation of total community RNA for amplification using the MessageAmp protocol

Amplification of mRNA in a sample via Messageamp is dependent on the presence of polyA tails on the mRNA transcripts. The polyA tail is used as a binding site for a primer sequence containing a polyT region and a T7 promoter sequence which is used to prime a reverse transcription reaction. The resulting single stranded cDNA is converted to double-stranded cDNA which is itself used as a template for the production of large numbers of copies of RNA molecules. The samples investigated here were from an environment where a large proportion of the population was likely to consist of bacteria and archaea, in addition to eukaryotes so an initial polyA tailing step was performed.

The PolyA tailing reaction was performed using *E. coli* PolyA polymerase (New England Biolabs). Reaction conditions are summarised in Table 4.2. The reaction was incubated at 37 ^{o}C for 30 min.

Table 4.2. Components of the Polyadenylation reaction.

Component	Quantity
10x reaction buffer	2 µl
10mM ATP	2 µl
PolyA polymerase Enzyme	1.6 µl (8 units)
RNasin RNAse inhibitor	0.5 µl
RNA sample	Volume equivalent to 150 ng of RNA
Water	To final volume if needed
Final Volume	20 µl

RNA from before and after the polyA tailing reaction was visualised using agarose gel electrophoresis. (Fig 4.1). This visual confirmation provided sufficient confidence to utilise the sample for RNA amplification. The RNA was cleaned up with the Qiagen RNeasy minelute cleanup kit, in addition to the cleanup step performed as part of the MessageAmp protocol, before further use.

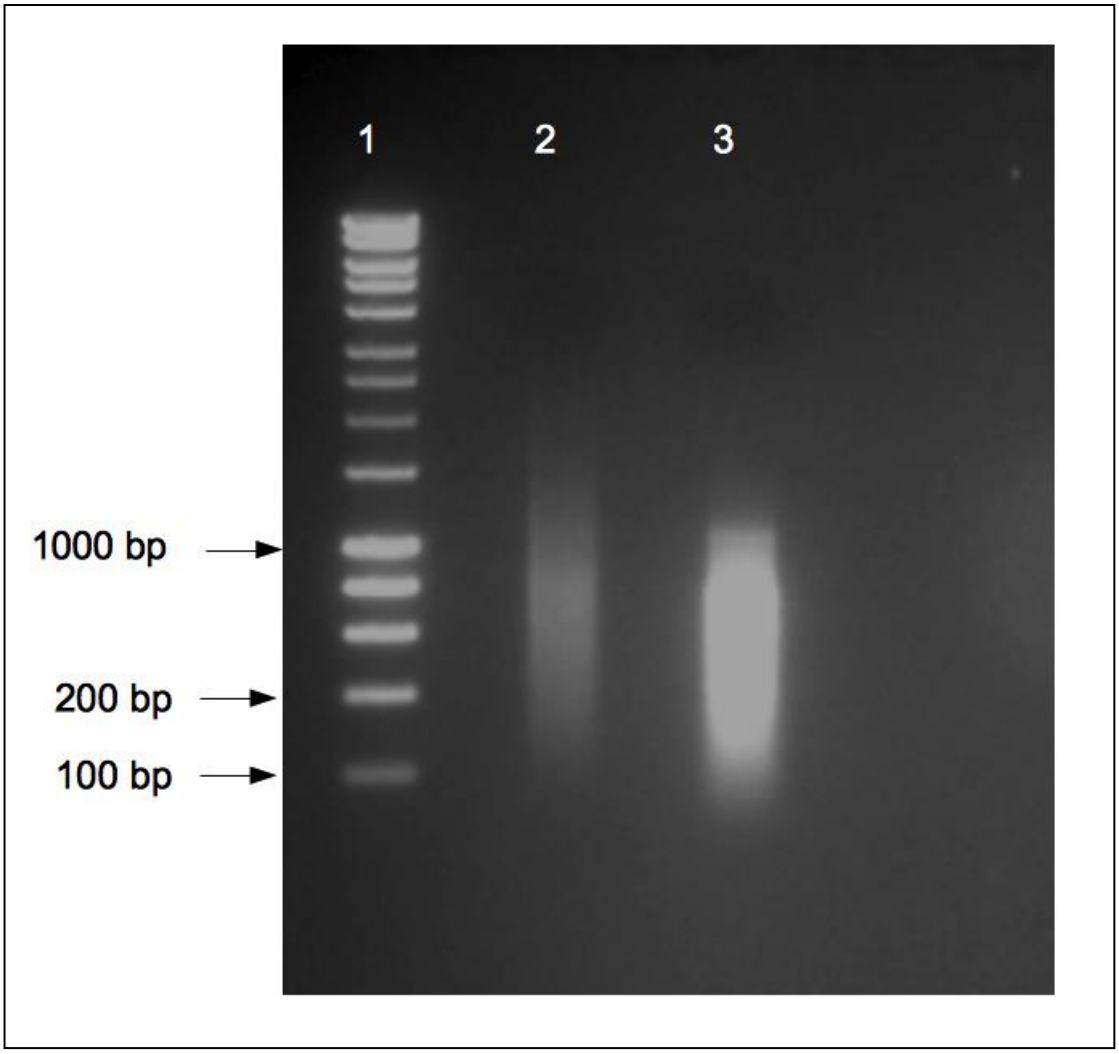


Fig 4.1 Profiles of the RNA samples before and after the polyadenylation reaction. Lane 1, size marker (Hyperladdder I, Bioline); lane 2, polyadenylated RNA; lane 3, untreated RNA. There is an apparent shift upwards of the profile of the RNA that had undergone polyadenylation of approximately 50 bp. The difference in brightness levels is simply due to a small aliquot of the initial sample being used for the reaction and becoming diluted into a large volume to make up the reaction mix.

4.2.3 MessageAmp amplification of polyA-tailed RNA.

Two RNA samples were processed in parallel using the MessageAmp kit according to the manufacturer's protocol. The two RNA samples consisted of 158 ng and 137 ng of starting material, which was towards the lower limit (100 ng) of the manufacturer's recommended starting quantity. On completion of the messageAmp protocol, the yields obtained were 10 µg and 9.1 µg. Maximum obtainable yield from 100 ng of RNA starting material estimated by the manufacturer, using high-quality control RNA, is 15 µg, and when using 1000 ng of starting material 176 µg. Extrapolating from this approximately linear relationship, the maximum yield when working with control 150 ng of RNA would be 15 times that value, which is 2.25 µg. The RNA yield data is provided by the manufacturer with the caveat that experimental RNA samples are expected to have widely differing yields and that RNA quality is a key determinant of the performance of this kit. Although a great deal of effort was expended to produce RNA of a high quality for amplification, given the unusual and challenging nature of the samples it was expected that maximum efficiency would not be achieved. The performance of the MessageAmp kit with this sample produced more than enough material for use in downstream applications and even though the amount of RNA was lower than theoretical maximum yield, this does appear to be a very robust method for amplifying small quantities of RNA prepared from the samples and with the methods used in this project.

4.2.4 Production of a dscDNA library from amplified RNA.

1 µg of RNA from each of the aliquots of RNA produced via the messageAmp protocol was used to produce an equivalent amount of dscDNA. The dscDNA aliquots obtained were submitted to the Centre of Genomic Research for sequencing on the 454 platform.

4.2.5 Testing an rRNA subtraction method for RNA sequencing.

Production of the initial library above did not include an rRNA subtraction step. This allowed for a faster library production workflow, with fewer steps and less handling of the sample, alleviating concerns about contamination of degradation-prone RNA becoming likely during a longer process. On the other hand, the fact that a polyA tailing step needs to be employed as part of the MessageAmp protocol when working with bacterial RNA means that all the RNA in the sample will have been polyadenylated including rRNA sequences. Since only mRNA is of interest here, it was decided to investigate the use of a rRNA-subtraction method to remove unwanted rRNA sequences prior to polyA tailing of the sample and subsequent amplification. The method chosen was a custom protocol described by Stewart *et al*

(2010), rather than a commercial kit-based method so that the probes used for rRNA subtraction would be sample-specific. The steps involved in metatranscriptome library preparation with and without rRNA subtraction are summarised in Fig 4.2.

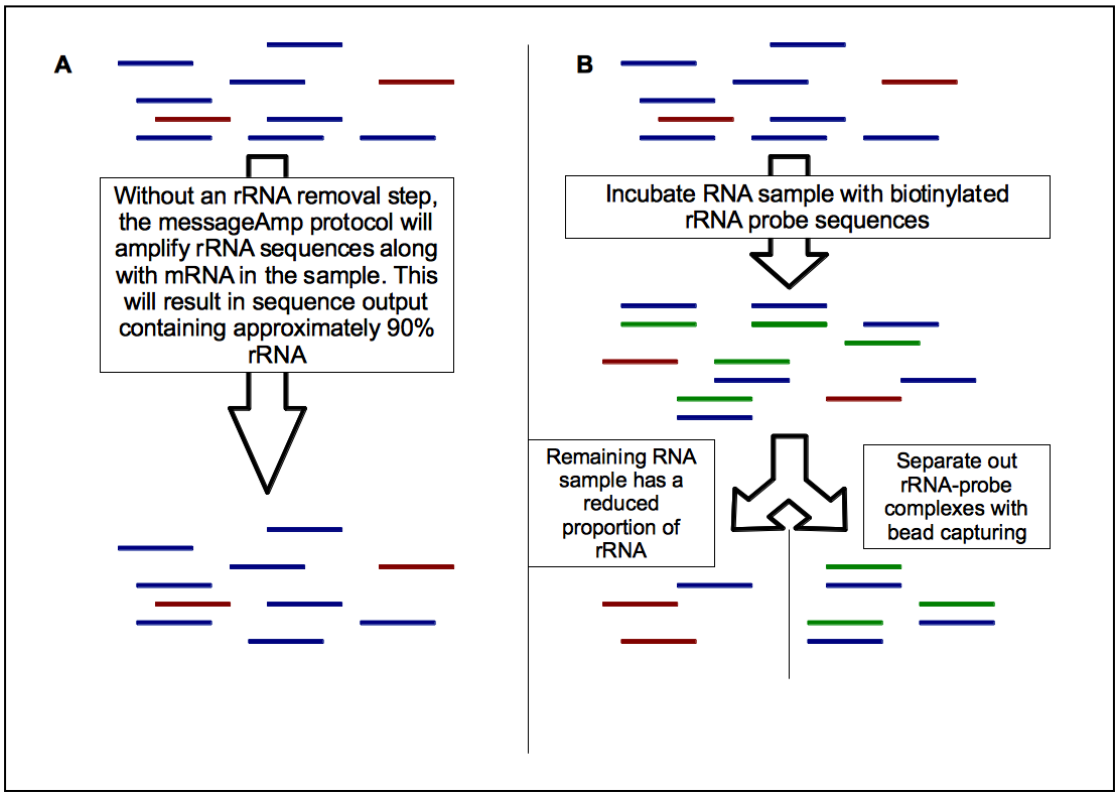


Figure 4.2 Comparison of total community RNA processing: **A,** without rRNA subtraction and **B,** with subtraction using the method of Stewart *et al*. (2010). Other rRNA subtraction methods exist, using probes or other enzymatic means to selectively remove rRNA sequences but all are conceptually similar in that they are designed to remove as great a proportion of rRNA as possible in order to yield a final product comprising mostly mRNA sequences for those applications focussing on active transcripts and expressed genes.

4.2.5.1 Method

Fig 4.3 illustrates the workflow for producing custom rRNA probes according to the method of Stewart et al. (2010). Briefly, this method requires DNA and RNA to be co- extracted from the same sample. The DNA is then used to produce biotinylated probe sequences that are in turn used to hybridise to the rRNA sequences in the total community rRNA sample and retrieve them using streptavidin-coated magnetic beads and magnetic separation.

Universal 16S and 23S rRNA gene primers, where one primer from each set contains a T7 promoter site, were used to amplify the almost full-length of the ribosomal genes from total community DNA. This amplification was carried out using the Herculase polymerase (Section 2.3.3) and the Eub16S and Eub23S primer sets (Table 2.2). This yields amplicons corresponding to the 16S and 23S rRNA sequences in the sample, with a T7 promoter site incorporated. This enables the amplicons to be used as templates for the MEGAscript high yield transcription kit, which allows the production of biotin-labelled anti-sense RNA probes, the sense being determined by how the T7 promoter site was incorporated during the initial PCR. The reaction conditions for probe production are listed in Table 4.3. The reaction was carried out at 37^{o}C for 6 h.

Table 4.3 Megascript reaction components.

Component	**Quantity**
PCR amplicons of rRNA genes	1 µl (approx. 500ng)
ATP (75 mM)	2 µl
GTP (75 mM)	2 µl
CTP (75 mM)	1.5 µl
UTP (75 mM)	1.5 µl
Biotin-11-CTP (10mM)	3.75 µl
Biotin-16-UTP (10mM)	3.75 µl
10x Reaction Buffer	2 µl
T7 RNA polymerase	0.5 µl
RNasin RNAse inhibitor	0.5 µl

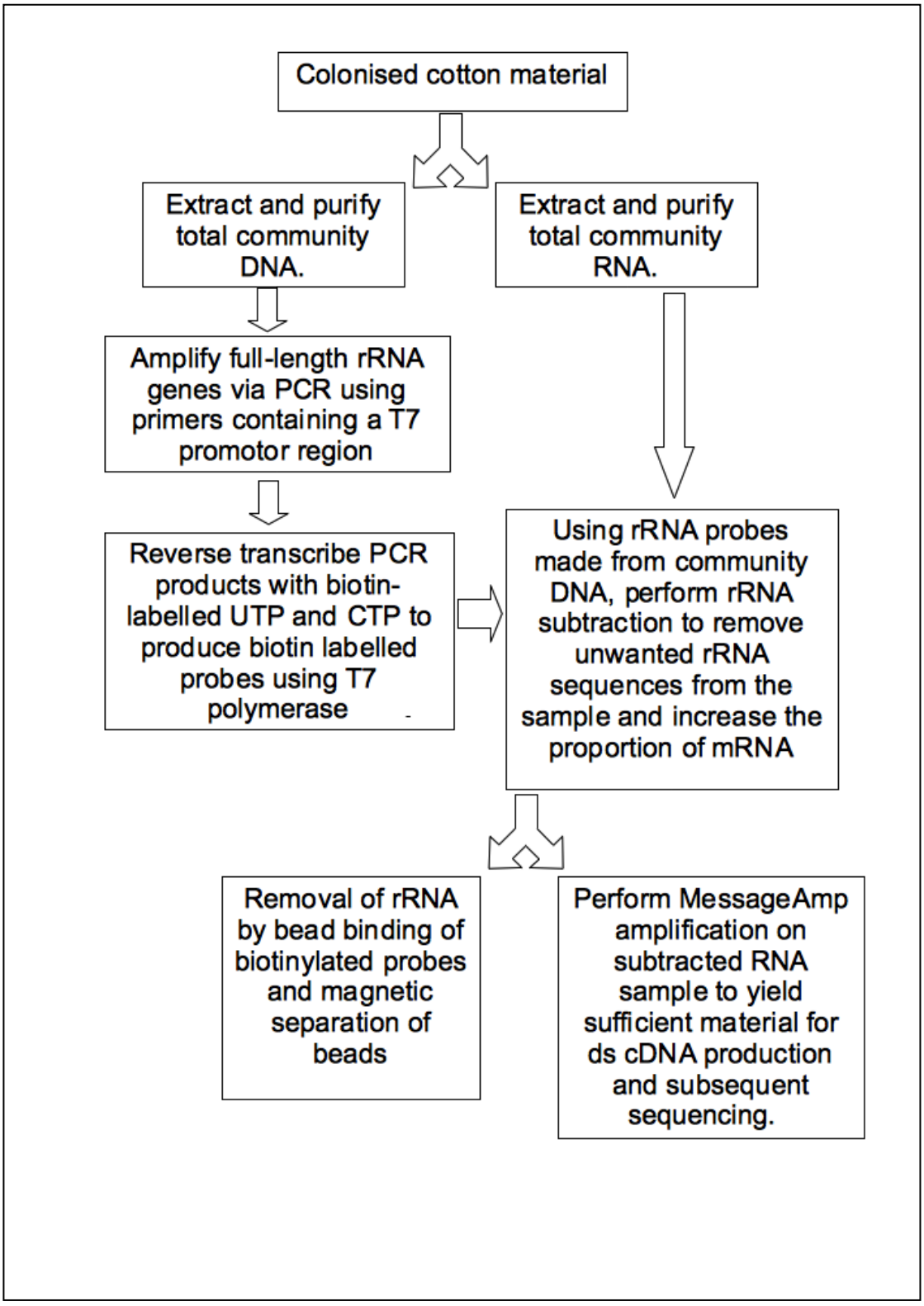


Figure 4.3: Summary of the rRNA probe-based subtractive hybridisation method described by Stewart *et al*. (2010).

Biotinylated probes can then be incubated with the RNA sample to allow hybridisation to the rRNA sequences. which can the be removed via streptavidin coated magnetic beads. The hybridised RNA sample is transferred to beads aliquoted into microcentrifuge tubes placed in a magnetic rack. The hybridised rRNA probe complexes are captured by the beads which are in turn captured by the magnetic rack and remain adhered to the side of the tube. After a short incubation of a few min at room temperature, the remaining RNA sample can be removed by careful pipetting, now subtracted of its rRNA component.

4.2.5.2 Results.

The rRNA-subtracted library was to be made from the same RNA stocks as the initial non –subtracted library. DNA for PCR template material was taken from freezer stocks that contained DNA extracted at the same time as the RNA stock was produced. The DNA had been maintained at -80^{o}C prior to use. PCR amplification was carried out to generate universal 16S and 23S rRNA gene sequences from the metagenomic DNA. 16S rRNA gene products of the expected size were generated in sufficient quantity. 16S rRNA probes were successfully made from the 16S rRNA gene amplicons using the MEGAscript kit.

Amplification of 23S ribosomal gene sequences from the sample DNA did not yield a single product of the expected size and therefore was not able to generate amplicons that could be used to produce rRNA probe sequences. Amplification of a discrete product band of the expected size for the 23S rRNA gene primers was achieved using *E. coli* genomic DNA. Repeated attempts to amplify a 23S rRNA gene product from the environmental sample were however never successful.

The 16S rRNA probes were stored at -80^{o}C but ultimately were not used for rRNA subtraction as it was decided that the lack of accompanying 23S probes rendered this whole approach somewhat ineffective. Alternative methods of rRNA subtraction were available and therefore it was decided to employ one method in its entirety. The 16S rRNA probes could have been used to achieve some rRNA removal in addition to another method but introducing extra processing steps always risks damage to the RNA sample so use of these sample specific probes as an approach to ribosomal RNA depletion was abandoned.

4.2.6 Results of 454 sequencing a cDNA library produced from a polyadenylated and amplified RNA sample.

The pyrosequencing performed on the non rRNA-subtracted library produced a very poor output with over 90% of the reads <70bp in length. The expert advice from the Centre for Genomic Research who performed the sequencing was that this

is indicative of poor sample quality, as similar output was occasionally generated from such cDNA samples and may be related to over-amplified cDNA during library production caused by PCR amplification. Some cDNA production kits can lead to incorporation of adaptor sequences which contain homopolymer sequences, which are a particular problem for the 454 pyrosequencing platform and frequently lead to large scale sequencing errors. In this case, the protocol used included the MessageAmp kit and the Just cDNA double-stranded cDNA kit, neither of which used a PCR amplification step to produce cDNA. The MessageAmp kit uses an *in vitro* reaction to produce large amounts of RNA from a cDNA intermediary but this amplification step does not require the use of adaptor sequences which can lead to the creation of homopolymer sections on the ends of reads. The problem in this case was almost certainly the result of the polyA tailing step performed prior to messageAmp amplification. The addition of polyA tails to all of the RNA molecules in the sample would have been carried through the messageAmp and subsequent dscDNA creation process and therefore the cDNA for sequencing would have contained long poly adenine tails at either the 3' or 5' end. The effect of the polyA tailing step was not appreciated before submission of the sample and the technicians who performed the sequencing were not fully aware of how the sample had been prepared. The end result of this was a catastrophic sequencing run, with the long polyA tails conflicting severely with the homopolymer-sensitive detection apparatus of the 454 machine, and a dataset that was effectively useless.

4.2.7 Re-design of metatranscriptome library preparation for sequencing on the Illumina MiSeq platform.

After the poor results from the initial metatranscriptome sequencing effort and from the attempt to produce a second library with rRNA depletion, a new library preparation protocol was developed with the intention of circumventing both of these problems.

Firstly, the new library would be sequenced on the Illumina MiSeq platform which is not homopolymer sensitive, so the polyA tailing step necessary for MessageAmp amplification would not be problematic. At the time of sequencing (August 2012), the MiSeq platform was limited to 150x150 paired end sequencing and so it was decided that incorporation of a step to remove the polyA tails would be a useful addition to save wasted sequencing of those areas of the RNA and maximise sequencing of potentially interesting protein-coding sections. Removal of polyA tails has in fact been used successfully to enable the pyrosequencing of bacterial community mRNA amplified with the MessageAmp method (Frias-Lopez *et*

al., 2008) and with this step added into the library preparation workflow, this work could have been carried out on the 454 platform. However, the MiSeq platform yields many more reads per sequence run, although at a reduced read length, and this factor in addition to a “once-bitten-twice-shy” sentiment rendered the switch from 454 to Illumina sequencing a sensible choice.

Secondly, an rRNA subtraction method would be employed which would not rely on a commercial-based kit or custom made probes. The Terminator 5’-phosphate dependant Exonuclease enzyme from Epicentre is active against any RNA sequence with a 5´-monophosphate end, and is an excellent choice for rRNA subtraction in a sample harbouring unknown species which commercial kits may not be completely effective at processing. It is also relatively cheap and easy to use.

4.2.8 Assessment of the effectiveness of Terminator 5’-phosphate dependant Exonuclease for rRNA subtraction.

A community RNA aliquot was treated with Terminator enzyme. The reaction was carried out using the reagents listed in Table 4.4. The manufacturer provides two reaction buffers, a recommended buffer and alternative buffer. The recommended buffer confers a higher activity on the enzyme but at a slight risk of possible non-specific activity against some mRNA species in the sample. The alternative buffer has no risk of no-specific degradation but confers a reduced amount of activity on the enzyme. Given the nature of the sample, it was decided to use the alternative buffer and avoid any potential mRNA degradation. The reaction was incubated at 42 ^{o}C for 30 min and was inactivated by the addition of 1 µl of 100mM EDTA, and the RNA purified using the RNeasy MinElute cleanup kit (Qiagen). RNA was stored at -80^{o}C prior to use.

Table 4.4 Terminator Exonuclease reaction components.

Component	Quantity
Terminator 5x reaction buffer B	2 µl
RNasin RNAse inhibitor	0.5 µl
Terminator Exonuclease	1 µl
RNA sample	X µl – as much as is required up to 10µg
RNAse-free water (if necessary)	Up to 20 µl

RNA from before and after the Terminator digestion step was assessed via qPCR (protocols carried out as described in detail in section 2.5), using universal bacterial 16S rRNA gene primers (Table 2.2) to enumerate copy numbers of SSU ribosomal sequences and to therefore determine the proportion of rRNA in the sample that had been digested by the Terminator enzyme.

Copy numbers for 16S rRNA in the RNA sample, measured as reverse transcribed cDNA, before and after treatment were determined and the copy number of the treated sample was expressed as a percentage of the copy number of the untreated sample. The treated sample had 17.5% (+/- Standard deviation of 1.4) of the 16S rRNA compared to the untreated sample, the quantity of rRNA of which was normalised to 100% (+/- standard deviation of 9.5). Therefore, 82.5% of the rRNA in the sample was removed, assuming equivalence for 23S rRNA. Although removal was not complete, the reduction was reasonable and greatly improved the ratio of mRNA to rRNA in the sample so that the final sequencing output would contain a greater proportion of protein coding sequence data.

4.2.9 Incorporating PolyA tail removal into dscDNA library production from community RNA.

Using the technique of Frias-Lopez *et al*. (2008), amplification of RNA with the MessageAmp protocol was carried out replacing the poly T/T7 promoter site containing primer supplied with the kit with a custom primer. The custom primer includes a poly T and T7 section as the original but also a sequence corresponding to the recognition site for the BpmI restriction endonuclease. BpmI will cut DNA when it recognises a CTGGAG$[N]_{16}$ region. Therefore, if the BpmI recognition site is incorporated at the beginning of the messageAmp protocol and subsequently carried through the entire process until final amplified RNA and dscDNA production, the final dscDNA product can be treated with BpmI to remove the polyA tail prior to sequencing. This neatly removes a large amount of the homopolymeric material resulting from messageAmp amplification. A schematic summary of how this process is carried out is presented in Fig 4.5.

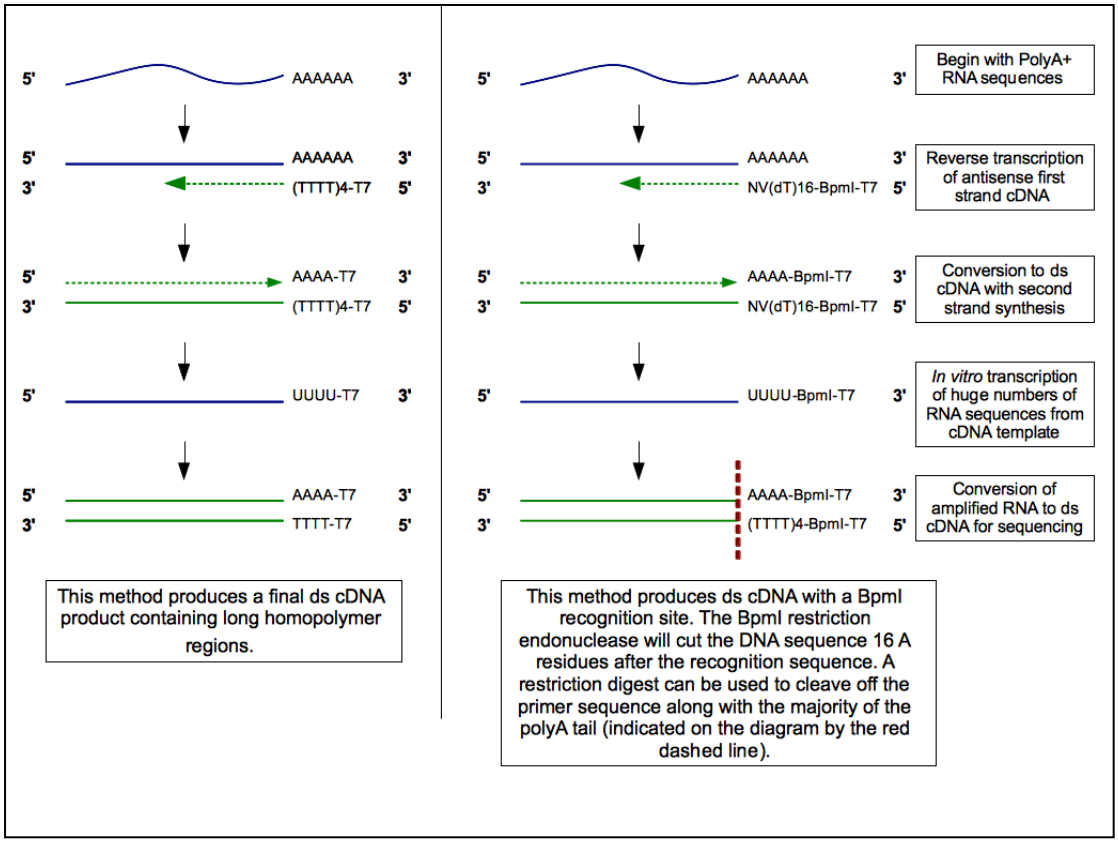


Figure 4.5 An overview of the MessageAmp RNA amplification process showing the standard protocol and a modified version which allows incorporation of the BpmI recognition site and subsequent removal of polyA tails with a restriction digest.

4.2.10 Production of Terminator-treated, polyA tail removed dscDNA for metatranscriptome sequencing.

Total community RNA was rRNA subtracted using the Terminator enzyme protocol,polyadenylated and then used as input material for the messageAmp protocol. A total of 23 µg of amplified RNA was generated and used as template for reverse transcription to dscDNA which was digested with BpmI to remove polyA tails. The BpmI digestion was performed in 50 µl volumes with reagents as listed in Table 4.5 and incubation of the reaction was for 37^{o}C for 1 h. Reaction components were 40 µl of cDNA in dds$H_2$0, 5 µl NEB Buffer 3 and 5µl BpmI enzyme.

The BpmI digested cDNA was cleaned up with the Bioline PCR and gel kit and quantified on the qubit to ensure sufficient source material for library preparation and sequencing. The cDNA was also visualised via agarose gel electrophoresis to check integrity.

The final dscDNA production step in this amended workflow was not very efficient. When the messageAmp protocol was first used to generate amplified RNA for cDNA production, two aliquots of cDNA comprising 1.2 µg in total were produced from 1 µg of starting RNA, which represents an efficiency of 60%. After incorporation of the BpmI recognition site into the RNA amplification process it was necessary to perform and pool together twelve cDNA aliquots produced from the same quantity of starting material. Why the cDNA synthesis step become so highly inefficient is unclear.

4.2.11 Production of a metatranscriptome: Concluding remarks.

Amplification of RNA with the MessageAmp kit absolutely requires polyA regions for priming of the initial reverse transcription step and so a polyadenylation of the environmental RNA sample was performed accordingly. The presence of long polyA regions introduced onto the end of the RNA sequences here was the most likely explanation for the failure of the initial attempt to sequence a metatranscriptome using 454 pyrosequencing. Subsequent to this, the method for producing a dscDNA library from RNA starting material was heavily altered to include removal of both polyA tails by restriction endonuclease digestion and excess rRNA sequences. These improvements produced a better quality library, with depletedrRNA sequences and a greater enrichment of mRNA reads. The cDNA sample was submitted to the Centre for Genomic Research (CGR) at Liverpool for sequencing. The quality assessment and bioinformatic analysis of this dataset is presented in Chapter 5.

4.3 Extraction of high molecular weight DNA for production of a fosmid library.

To generate a fosmid library for screening of intact genes, the CopyControl Fosmid library Production kit from Epicentre was used. This kit is capable of yielding 80,000 fosmid-containing *E coli* clones at maximum efficiency (Epicentre). Use of this protocol requires high molecular weight DNA in the size range of approximately 30-50 kb.

4.3.1 Initial testing of extraction methods.

While the Griffiths *et al*. (2000) method (section 2.2.2) was perfectly adequate for generating clean RNA and DNA suitable for PCR amplification and pyrosequencing from the colonised cotton samples it was found to be unsuitable for production of HMW DNA. DNA produced by this method was heavily sheared as a result of the bead-beating step in the method. Figure 4.6 below illustrates the appearance of DNA extractions obtained from colonised cotton from Esthwaite water using this method. The size range was < 6000 bp. A much gentler method was therefore required to produce HMW DNA and avoiding multiple pipetting steps during the extraction process. The method of Neufeld *et al*. (2007) was selected, as it does not include mechanical disruption of the sample.

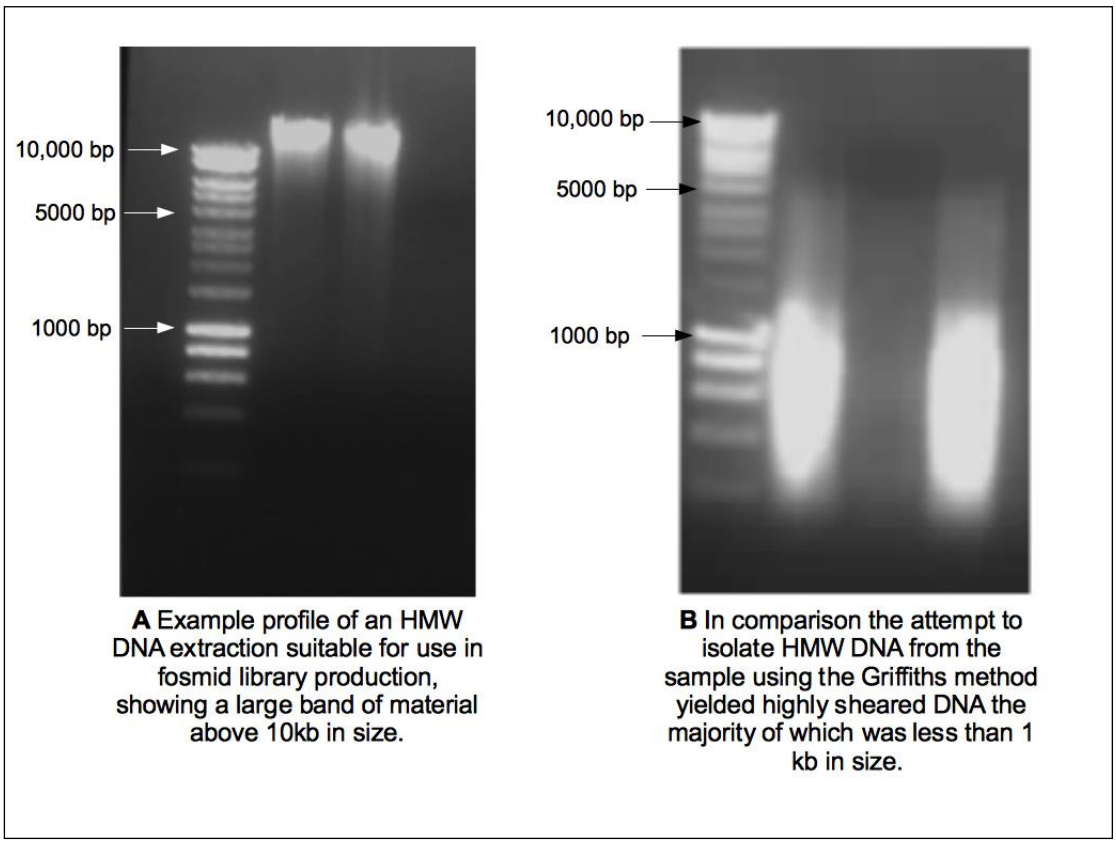


Fig 4.6 Comparison of DNA suitable for fosmid library production with a DNA extraction produced from the Esthwaite colonised cotton bait sample demonstrating that intact HMW DNA from this source required gentle lysis.

Neufeld *et al*. (2007) used a gentle extraction method to produce metagenomic DNA from sterivex filters through which seawater had been filtered to collect microbial cells. The method comprises gentle washing steps and phenol:chloroform extraction using phase-lock tubes that contain a gel which, during centrifugation, will stratify between the organic and aqueous phase of the mixture and separate the two layers. The organic solvents and various impurities dissolved within them are trapped at the bottom of the tube by the gel. The top aqueous layer can simply be decanted off, avoiding all pipetting of the sample at this stage of the process.

The extractions were carried out as follows, with buffer recipes as described in section 2.8.1. Approximately 5g of colonised cotton was placed in falcon tubes and 5 ml of SET buffer and 10 µg of fresh lysozyme solution were added. The tubes were incubated at 37°C for 30 min with gentle shaking. 500 µl of 10% SDS and 55µl of proteinase K solution were added and the incubation was continued for 2 h at 55°C with gentle shaking (20 rpm). The lysate was removed by decanting into phase-lock tubes and the colonised cotton was washed with an additional 1ml of SET buffer and subjected to gentle shaking. The rinse buffer was decanted and added to the lysate. An equal volume of phenol:chloroform was added and the tubes were centrifuged for 5 min at 4000 x g. The phenol:chloroform extraction was repeated and a final chloroform extraction carried out with an equivalent volume and centrifugation as before. Finally, the aqueous phase was transferred to a centrifuge tube and 5µl of 20 µg ml^{-1} glycogen, 0.25 volumes of 7.5 M ammonium acetate and 2 volumes of absolute ethanol were added. The tubes were then incubated overnight at -20 °C, and the nucleic acids were then pelleted by centrifugation at 48 000 x g. The pellet was air dried, washed with 70% ethanol, and resuspended in SDD H_20.

4.3.2 Results

DNA was successfully obtained, and in high quantity, with amounts varying between 1.5 and 6 µg across various extractions. Visualisation via gel electrophoresis demonstrated that HMW DNA was obtained although the overall size range of the DNA included a great deal of material < 25 kb. The actual amount of DNA suitable for use in the CopyControl protocol was therefore effectively much lower but the kit calls for 0.25 µg of HMW DNA at the key ligation step and this extraction method appeared to have a high enough yield so it was possible to be confident that size selection from multiple cotton bait extractions would produce the quantity of DNA required for the intended application. A size selection step was then carried out to separate the HMW DNA from the smaller fragments.

It was noted that the DNA pellet obtained after ethanol precipitation was a rust colour, which is most likely an indication of contamination by humic substances from the lake sediment. The DNA extraction method used here was not designed with sediment or soil samples in mind and as a consequence the DNA obtained was heavily contaminated. As a size selection step was needed it was thought that the electrophoresis and subsequent gel extraction would be effective at cleaning up the sample and provide clean DNA suitable for use in fosmid library production.

4.3.2 Size Selection.

A size selection step was performed according to instruction provided by the manufacturer using pulse-field gel electrophoresis (PFGE) with low melting point (LMP) agarose. In order to keep the DNA undamaged for production of fosmid clones, it must not be exposed to UV light and it must be kept free of Ethidium Bromide contamination. The kit protocol therefore suggests a method where size markers are run in the outside lane of a gel and the actual sample run in central lanes. After electrophoresis, the outside marker lanes can be cut off from the main part of the gel and post-stained. Visualisation of the stained size marker lanes allows the point to where the bands of the desired size have migrated to be marked. The stained marker sections are then lined up with the unstained part of the gel, so the marked position of the size fragments can be used to pinpoint the position of the unstained sample DNA. The section of the gel where this DNA is located can be cut out and gel extraction performed to yield DNA of the desired size. This process is illustrated in Fig. 4.7.

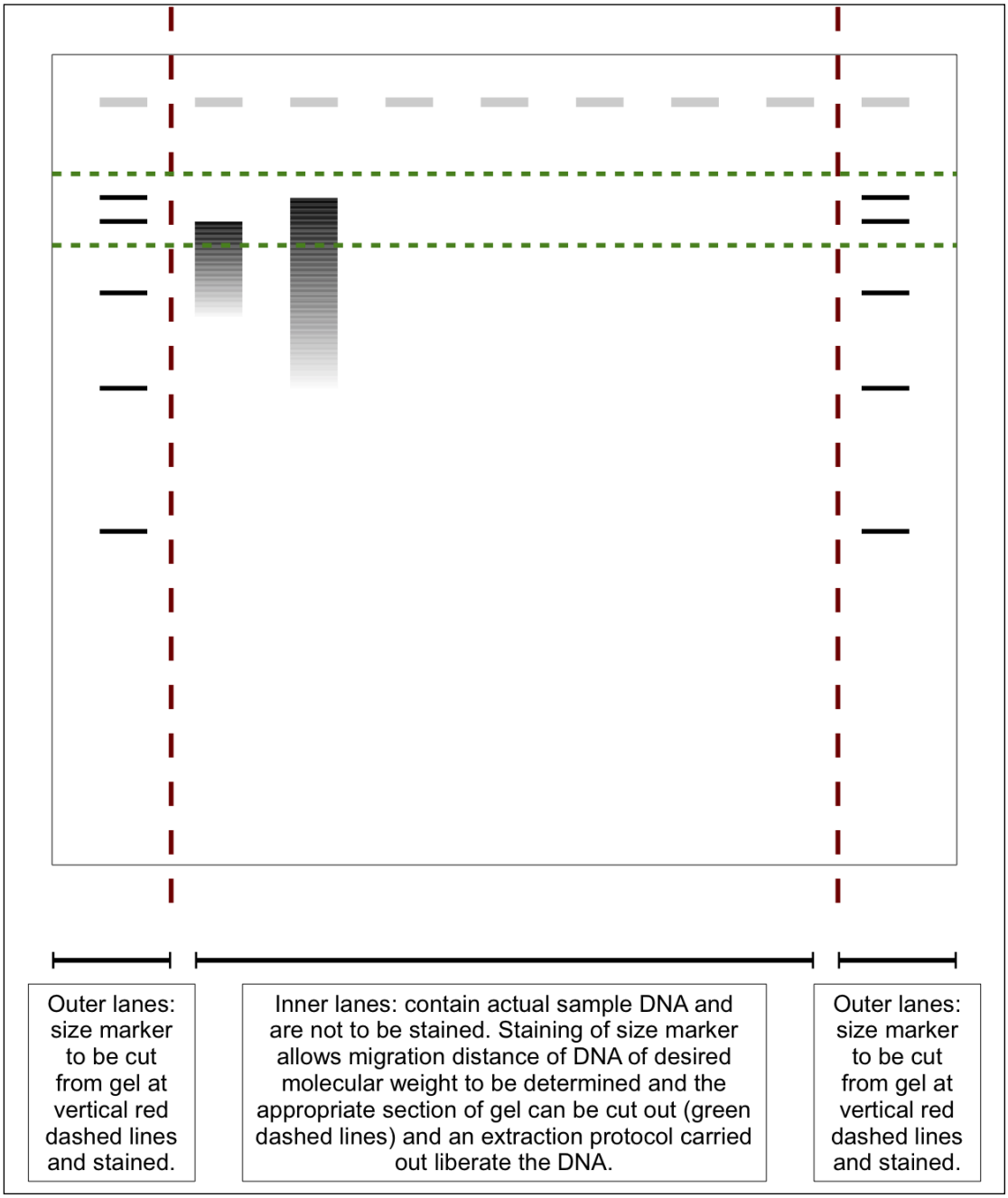


Figure 4.7 Processing of PFGE gels for the removal of HMW DNA in the size range of 30-40 kb.

Briefly, the sections of the agarose containing HMW DNA were cut into pieces of approximately 500 mg each, to yield 500 µl of molten agarose, and the pieces were placed into separate microfuge tubes. The LMP agarose was melted at 70°C for approximately 15 min until it was fully liquefied then allowed to equilibrate to 45 °C. 5 µl of 50x Gelase buffer pre-warmed to 45 °C was added to obtain a 1x concentration. 5 µl (equivalent to 5 Units) of Gelase enzyme was added and the tube was incubated at 45 °C for 1 h. The enzyme was inactivated at 70 °C for 10 min. Tubes were chilled on ice for 5 min and then centrifuged at maximum speed in a microfuge for 20 min. This step pellets various insoluble oligosaccharides and removes unwanted material from the DNA. The top 90% of the supernatant was removed to a new tube and precipitated with 2 volumes of 100% ethanol and ¼ volume of 3 M sodium acetate. The tubes were centrifuged in a microfuge at full speed for 20 min to pellet the DNA. The pellet was rinsed twice with 70% ethanol, and ethanol removed by vacuum centrifugation at 40 °C for 15 min.

Agarose may have been associated with the DNA pellet, and could not be visually differentiated easily. A second digestion tended to cause loss of DNA. At this stage it was also difficult to verify the size of the DNA as it was necessary to conserve all of the material for subsequent steps. The effects of extra handling that second digestion would have inflicted upon the integrity of the HMW DNA could therefore not be assessed fully. The agarose was melted at 70°C for approximately 15 min and then equilibrated to 45 °C

Qubit quantification identified very low amounts purified from the agarose gel, at concentrations of < 1 ng. The extremely low yield obtained may have been due to experimental error or poor quality DNA material. Repeated extractions and attempts at size selection with this protocol were performed with similar yields achieved each time. It was concluded that the sample itself was the problem.

4.3.3 Troubleshooting and method development for HMW DNA size selection.

There were a finite number of cotton baits harvested from Esthwaite water available for HMW DNA extraction. After the initial failed attempt at size selection and a subsequent repeat of the experiment, the remaining cotton baits therefore become a precious resource. For the purposes of testing alternative workflows and designing new methods it was decided to produce some colonised string using a carboy of landfill leachate (microcosm 1 in chapter three). The microbial community in this microcosm colonised cotton baits rapidly, forming a biofilm, and extractions from cotton baits placed in this microcosm yielded large quantities of DNA after only two weeks of incubation; this characteristic meant that samples for extraction could

be generated very quickly for use in experiments. In addition, the leachate was by nature “dirty” containing various contaminants that required a thorough cleanup to yield DNA suitable for molecular biology work (McDonald *et al*., 2008, McDonald *et al*., 2010). These characteristics made it a suitable surrogate for the Esthwaite cotton samples, and DNA extractions for method development of HMW DNA extraction, cleanup and size selection. Cotton baits from this carboy of landfill leachate were therefore used as the source of DNA in the troubleshooting experiments.

In fact, the DNA that was extracted from the landfill leachate-colonised cotton baits was “cleaner” than the lake DNA extractions. The pellet obtained after extraction, purification and precipitation did not possess the same rust coloured appearance. The DNA, when visualised on a gel, appeared to have suffered less degradation as a large clump of DNA at or above 10 kb was visible without a long smear of lower molecular weight material. This is illustrated in Fig. 4.8, by comparison with Fig. 4.6. It would seem that the biofilm on the leachate-colonised cotton is much more amenable to the extraction process. This could be an indication that the material removed from the lake had not stored well, and although it had been kept at -80°C until use it is possible the freeze-thaw process itself led to degradation of the biofilm DNA. New cotton baits were prepared for introduction into Esthwaite water so that fresh material might be available at a later date.

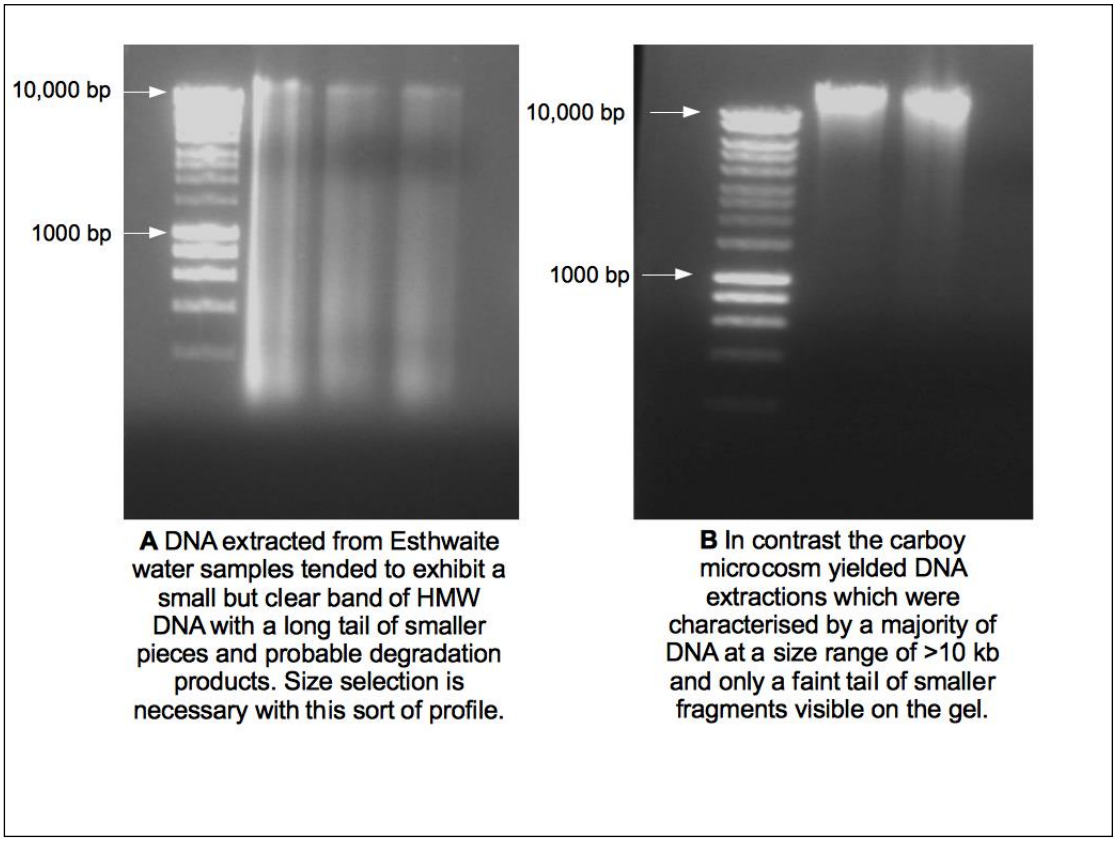


Figure 4.8 A comparison of the agarose gel profiles of DNA extracted from Esthwaite water and leachate carboy microcosm cotton baits, produced using the same method, demonstrating the difficulty of extracting high quality material from the lake samples, the DNA of which appeared to be highly susceptible to degradation.

4.3.3.1 High Gelase activity protocol.

The protocol for extraction of DNA from agarose gels provided with the CopyControl kit literature differs from the method described in the literature that Epicentre provide when Gelase enzyme and buffer are purchased as a separate component. It was decided to investigate the protocols described in the dedicated Gelase instructions to determine if an extraction could be achieved which avoided the apparent carry over of agarose, specifically using the described “High activity protocol”. Gel slices were cut and placed into tubes as before. 3 µl of 1x Gelase Buffer per mg of gel were added to each tube. Each tube was incubated at room temperature for 1 h and the buffer was then removed and discarded. The agarose pieces were incubated at 70°C for approximately 15 min until it was fully liquefied then allowed to equilibrate to 45 °C. Insoluble oligosaccharides were removed as for the protocol described in the kit. DNA was also concentrated by precipitation as per this protocol.

4.3.3.2 β-agarase digestion.

Another enzyme able to digest agarose was tested to see if any improvement in performance could be observed. β-agarase I was purchased from New England Biolabs and enzymatic digestion of agarose and isolation of DNA were carried out according to the recommendations of the manufacturer. Agarose pieces of approximately 500 mg were cut, placed into eppendorf tubes and 10x β-agarase I buffer was added to 1x concentration. Agarose was melted at 65 °C for 10 min, and cooled to 42 °C before addition of β-agarase enzyme at 1 unit per 100 µl of molten agarose, where it was assumed 100 mg of solid agarose yielded 100 µl when melted. The digestion was incubated 42 °C for 1 h., ¼ volume of 10 M ammonium acetate was added and the tube was chilled on ice for 15 min. To pellet insoluble carbohydrate material, the tube was centrifuged in a microfuge at full speed for 15 min. The supernatant containing DNA was decanted to a fresh Eppendorf tube and precipitated with 2 volumes of ethanol and ¼ volume 10M ammonium acetate. DNA was pelleted by centrifugation at full speed in a microfuge for 20 min, washed with 70% ethanol and centrifuged as before for 5 min.

4.3.3.3 Removal of HWM DNA from agarose using electroelution.

The fact that DNA liberated from gels via enzymatic methods was never able to generate fosmid libraries may have been due to inhibition caused by contamination from incomplete purification of the extracted DNA or resulting from the enzyme treatments and size selection steps, either from residual agarose, insoluble oligosaccharides associating with the DNA during size selection or buffer

components interfering with components of the fosmid kit. Another possible method of removing DNA from gel was electro-elution. Cutting out a small piece of gel, containing DNA, and applying an electric current to migrate the DNA out of the gel and into a sealed container of buffer allows the DNA to be transferred into an aqueous solution from where it can be purified and concentrated through precipitation.

Initially, a PFGE was performed as before to size-separate the HMW DNA and the size marker lanes were cut and stained in order to work out where the appropriately sized DNA fragments had migrated to in the other lanes to obtain gel segments containing DNA in the 25-50 kb range. The gel pieces were placed into sealed bags of boiled dialysis tubing containing 1x TAE buffer as for routine gel electrophoresis. The dialysis tubing bag was placed in a standard electrophoresis gel tank and maintained at a voltage of 40 V for 5 hours. Then dialysis bag was turned 180° and the current applied for one minute to ensure any DNA that had become bound to the walls of the dialysis tubing was liberated. The buffer within the tubing was then decanted into Eppendorf tubes, purified with a phenol:chloroform extraction with phase-lock tubes as previously used for DNA extraction.

4.3.3.4 Slow Soaking of agarose.

Another method explored for the liberation of DNA from agarose involved crushing or mashing small pieces of gel containing the DNA of the desired size range in a buffer solution and allowing the mixture to soak overnight at 37°C with gentle shaking. DNA was subsequently recovered with ethanol precipitation. Buffer composition was 300mM sodium aceate, 1 mM EDTA (pH8) SDS.

4.3.3.5 Troubleshooting and method development for HMW DNA size selection: concluding remarks.

Ultimately, enzyme digestion was not found to be an efficient means of liberating HMW DNA from LMP agarose gels in this case. DNA isolated from agarose was derived from the soaking method and electroelution methods. The actual quantity of DNA obtained from any individual digest was very low (typically in the region of 20-30 ng from a single digest) but pooling of dozens of extracts did generate sufficient material to perform an end-repair reaction with the CopyControl fosmid production kit and attempt the ligation and titration of clones to generate a fosmid library. No clones were produced by this method and the fosmid library production failed. Given that the DNA input into each step of the protocol for the CopyControl kit was in excess of the minimum recommended amount specified by

the manufacturer, it appeared that the purification of DNA from the agarose gels may have been introducing contamination which interfered with fosmid library production.

4.3.4 Avoiding size selection.

As size-selection using PFGE appeared to be causing problems, it was decided to attempt fosmid library production using non-size selected DNA, although such an approach is against the advice of the manufacturer of the CopyControl kit. The size selection step is required as there is a chance of generating chimeric clones when using non-size selected DNA with this kit, although the identification of expressed cellulase enzymes in chimeric clones would still be of value, although requiring more careful data analysis at a later stage. The CopyControl protocol includes an *in vitro* lambda packaging step which is naturally programmed to package DNA fragments that are the same length as its own genome, which is approximately 48 kb, but multiple pieces of DNA of <25 kb have a chance of being ligated together and therefore packaged by the phage vector.

A parallel workflow was carried out using the control DNA provided with the Copy Control kit to act as a positive control when undertaking this experiment so that poor performance could be ascribed specifically to the kit, experimental error, or to the sample itself.

The DNA was end-repaired according to the manual provided with the CopyControl Fosmid library Production kit. The CopyControl protocol was continued with the ligation step, with the maximum recommended amount of DNA (0.25 µg) used in the ligation reaction, and subsequent packaging of DNA and titering of clones to determine final titre of fosmid clones. No clones grew during titration of the clone preparations, so this method was therefore unsuccessful. DNA sample quality (size range, chemical contamination or a combination of both) was thought to be frustrating the fosmid production process. Consequently, it was decided to improve the extraction process to produce a cleaner starting sample of DNA.

4.3.5 Further Method development for the production of high quality HMW DNA.

In order to obtain cleaner starting material, extra steps were incorporated into the method of Neufeld *et al*. (2007) to tailor the process towards purification of DNA from a sample with high levels of humic substances. A CTAB buffer wash was used as for the Griffiths method which had proven to be effective at removing contamination from these samples but was used at a lower concentration (2% as opposed to 5% by Griffiths *et al*., 2000) as used by Porteous *et al*. (1997) for the extraction of DNA from soil. In addition, a low pH phenol:chloroform extraction was

used as Mettel *et al*. (2010) had demonstrated that purifying RNA at lower pH avoids co-extraction of contaminating material to a great extent.

Using fresh colonised cotton baits harvested after being left *in situ* at Esthwaite Water for 10 weeks the extraction method was followed as previously carried out for the first few steps of the protocol, comprising addition of lysozyme, incubation at 37°C, addition of proteinase K and SDS, and incubation at 55°C for 2 h. At this stage, crude lysate was heated to 68°C. 2% CTAB solution pre-heated to 68°C was added, the tube was mixed by gentle inversion and the mixture was incubated at this temperature for 10 min. Tubes containing the crude lysate were then centrifuged at 4000 rpm for 10 min to pellet insoluble materials. Two phenol:chloroform extractions were then performed in phase lock tubes as before. After the second extraction, DNA was harvested from the aqueous phase by PEG precipitation carried out overnight at 4°C in two volumes of 30% PEG solution (30% (w/v) PEG 6000 in 1.6 M NaCl). The resulting pellet was resuspended in a pH 5 citric acid/sodium citrate buffer solution. A phenol:chloroform extraction using acid phenol:chloroform (pH 4.5) was carried out in phase-lock tubes as before. Finally, the aqueous phase was transferred to a clean tube and 5ul of 20 µg ml^{-1} glycogen, 0.25 volumes of 7.5 M ammonium acetate and 2 volumes of absolute ethanol were added. The tubes were then incubated overnight at -20 °C, and the nucleic acids were then pelleted by centrifugation at 48 000 x g. The pellet obtained from this process had a clean, white appearance as opposed to the rust-coloured appearance observed in previous samples. The pellet was air dried, washed with 70% ethanol, and resuspended in 200 ul SDD H_20. DNA was obtained at quantities of approximately 6-10 µg per extraction with acceptable purities as determined from the $A_{260}/_{280}$ and $A_{260}/_{230}$ ratios which were > 1.8 and 1.7 respectively. The DNA obtained was used with the CopyControl kit as before but once again did not produce any clones.

A commercial company was used (Bio S&T inc, Montreal, Canada) to attempt a fosmid library production. They were provided with a total of 400 µg of this HMW DNA material from colonised cotton baits from Esthwaite Water and were able to produce 80, 000 clones. Exact details of the methodology employed was not released by the company. However, Bio S&T normally guarantee a yield of 100, 000 clones from their fosmid library production, with a starting sample of 200 µg of DNA. In this case, only 80, 000 clones were obtained even after a second amount of 200 µg of starting material was supplied and their feedback was that this was indeed a difficult DNA sample to process.

4.3.6 Fosmid library production: Discussion and Conclusions.

Production of DNA of satisfactory molecular weight and purity from the Esthwaite water cotton baits was a significant challenge. Utilisation of this DNA to produce a fosmid library was then only possible with specialist third-party support. The problems encountered in this process can be described on a basic level as stemming from attempting to carry out a challenging experiment with challenging starting material.

DNA isolated from Esthwaite water colonised cotton always exhibited a long smear when analysed with electrophoresis, indicating a great deal of degradation. In the main, extracts performed from landfill leachate cotton baits were always of a much improved profile, with larger quantities of intact DNA with a molecular weight above 10 kb. Since the extraction process itself was a constant, this leaves a few possible variables as the explanation for the poor quality DNA. It was damage, or shearing, of the DNA that was the problem here rather than quantity which was never an issue. Amounts of DNA obtained from different extractions of different cotton baits could vary but this is somewhat logical given the highly heterogenous nature of the samples which were likely to differ greatly in terms of level of colonisation from one to the next.

It is curious that an apparent discrepancy exists between the information from the CopyControl product manual and the literature provided with the Gelase enzyme preparation. The CopyControl kit instructs that Gelase buffer be used in the extraction of DNA from the gel used at the size selection step. However the Gelase manual specifically states that Gelase buffer can inhibit lambda packaging reactions and should be avoided when extracting DNA for this purpose. The CopyControl kit does in fact utilise a lambda vector and it was though that the use of gelase buffer might be introducing a contamination that inhibited this part of the protocol to some extent. However, avoiding use of the gelase buffer did not in anyway improve the performance of the CopyControl kit with this sample. The source of the problems can be assumed to have come from elsewhere.

Performing the protocol using the control DNA provided by the manufacturer allowed the performance of the components of the kit to be constantly monitored. With time, and repeated attempts at performing the protocol, the number of clones produced from the control reactions did drop which is probably attributable to aging reagents becoming deprecated after multiple freeze thaw cycles. When the quality of the output from the control reactions dropped, the kit was replaced to ensure good quality reagents were being used for the experimental samples and that problems in the workflow could be attributed to the only variable between the control and

experimental. The lack of fosmids and clones produced from the experimental samples can be absolutely stated not to have been due to poor reagents.

The remaining possibility is that the DNA extracts from both landfill leachate microcosm and lake introduced some level of contamination, which inhibited the reaction at some stage, or even multiple points. Bio S&T were ultimately able to produce fosmid-containing clones successfully but only by using a very large amount of DNA as starting material. The specifics of their methods were not released but the CopyControl kit calls for 0.25 µg of DNA at the ligation stage. It may be that vastly exceeding this specification can contribute to success when working with environmental DNA, suitably purified by the final multiple step process developed here. Progress was further hampered by the near impossibility of obtaining size-selected DNA from the PFGE gels. Extractions were only able to yield very low quantities of DNA from the gels. Again, it may be that starting out with huge quantities of DNA yields a reduced but workable quantity for end-repair, ligation and packaging. This placed some pressure on the requirement for multiple colonised cotton baits as a finite number were available and only so many HMW DNA extractions could be performed.

Ultimately, the lack of clones renders the issue of the poorer quality DNA extractions from the lake cotton samples a moot issue. Improved quality DNA from the carboy did not really improve local performance. Still, it is curious that cotton baits from the lake produced consistently poor quality DNA extractions; the molecular weight of a great amount of the extracted material exceeded 10 Kb but only a small amount of material was ever within the desired size range of 30 – 40 Kb. This may be explained by the resident biofilms being sensitive to freeze/thaw cyles to an extent that being frozen once for long term storage then unfrozen just prior to use was enough to deal damage. Alternatively, the community in the lake sediment was likely not equipped to deal with an oxygenated atmosphere or UV damage to DNA. A brief exposure to daylight and oxygen when removing the baits from the lake was unavoidable as the cotton was pulled up through the water column. Although transferred to dry ice almost instantly, oxygen and UV may have caused some damage to the residents of the biofilm and the genetic material they contained.

Chapter 5: Analysis of a metatranscriptomic dataset.

5.1 Background.

Metatranscriptomic sequencing was carried out on a cDNA library produced from a sample of RNA extracted from the biofilm colonising a cellulose bait that had been maintained in the anoxic sediment of Esthwaite water, a freshwater lake. Sequencing of the metatranscriptome of this community of microorganisms was performed in order to specifically study the expressed genes of the community residents. Utilising a cotton bait rich in crystalline cellulose was intended to act as an enrichment mechanism for specialist cellulose degrading organisms and therefore increase the incidence of gene sequences with roles in cellulose utilisation in the data; the hypothesis was that cellulolytic organisms in the sediment would be recruited to the crystalline cellulose material as they are the only member of the community who can actually use it as a carbon source. However no microorganism exists in isolation and a lake sediment will be home to a diverse array of species. The dataset obtained through metatranscriptome sequencing is as much an overview of a complex community as it is a resource for the discovery of specifically cellulolytic genes. Analysis of the large dataset was undertaken with both a view to establish general phylogenetic and functional characteristics for the data in addition to a more specific search for enzymes present in the dataset that might be linked to metabolism of the cellulose of the cotton baits. The chief goals of the metatranscriptomic survey were therefore:

1. To investigate the composition of the microbial community and discover which groups are particularly abundant and therefore likely to have key ecological functions in this environment.

2. To identify protein coding sequences in the data and utilise this information to investigate the profile of expressed genes in the environment

3. Search the data for gene sequences which are likely to have a function in cellulose breakdown.

5.2 Sequencing of an environmental metatranscriptome.

The method for development for the production of a cDNA library for sequencing of the cellulose bait biofilm metatranscriptome was discussed at length in chapter 4 but a brief overview of the process from start to finish is presented here. Total community RNA was extracted from the cotton baits using the method of Griffiths *et al*., (2000). Excess rRNA was removed from the sample using the

Terminator 5'monphosphate-dependant exonuclease (Epicentre). The rRNA depleted sample was polyadenylated to add polyA tails to all remaining rRNA sequences for compatibility with the MessageAmp Kit (Ambion). The polyadenylated RNA sample was processed with the MessageAmp kit according to the manufacturer's protocol but using a custom primer sequence to incorporate a recognition site for the BpmI restriction endonuclease. The amplified RNA sample was converted to dscDNA with the Superscript III kit (Invitrogen/Life Technologies) used for first strand synthesis and the RevertAid Premium Double-Stranded cDNA Synthesis kit (Thermo-Fischer) used for second-strand synthesis. Finally, the dscDNA sample was treated with BpmI to remove polyA tails before sequencing. The BpmI-treated dscDNA sample was passed to the CGR for sequencing. CGR staff performed sample size selection, library prep and sequencing of the cDNA library on the Illumina Miseq platform, generating paired-end reads of 150x150 bp.

5.3. Analysis of the dataset.

Before undertaking analysis such as taxonomic classification and gene prediction, the raw reads were extensively processed to remove artificial duplicates and reads of low quality using the Prinseq pipeline (Schmieder & Edwards, 2011). Reads were quality filtered or trimmed by assessing the phred quality scores of the base calls (Kunin *et al*., 2008) and other criteria. A Quality score of 20 represents a 1% chance of the base being miscalled and was selected here as a cutoff point for filtering and trimming of sequences. Removal of rRNA reads from the datset was also carried out using the Ribopicker tool (Schmieder *et al*. 2012). Assembly of metatranscriptomic data was attempted but did not yield satisfactory results; small contigs were obtained and these were few in number. As an alternative, paired end sequences with overlapping sections were combined into longer reads of 200-300 bp.

The dataset was analysed using the MEGAN Metagenome Analyser software (Huson *et al*., 2011) and the MG-RAST webserver which automatically annotates metagenomic datasets (Meyer *et al*., 2008). These resources provide a functional and phylogenetic overview of datasets. They both also enable classification of sequences to SEED subsytems (Overbeek *et al*. 2005) and KEGG Orthologies (Kanehisa *et al*., 2008). The SEED subsystems are defined as sets of "functional roles that together implement a specific biological process or structural complex" (Overbeek *et al*. 2005) and assignment of a read to a subsystem indicates a specific role for the protein it represents. KEGG Orthology (KO) assignment of reads assigns a read as having a function related to a specific pathway. Investigation of which are

the abundant KO assignments for a particular dataset can be an informative insight into the community from which it is derived. Use of SEED subsystem and KO classification fro metatranscriptomic reads specifically identifies active pathways and functions being carried out by the metagenome.

A more detailed investigation of protein coding sequences was carried out by searching predicted ORFs against the Pfam database of Hidden Markov Models (HMMs) (Finn *et al*., 2010) representing families of proteins with related activity.

5.4 Sequencing output, quality control and data pre-processing.

The raw data produced from the sequencing run consisted of approximately 7 million paired-end sequences of 150 bp, in the format of two files of corresponding sequences pairs. Prinseq (Schmieder & Edwards, 2011) was used to perform an initial quality check of the data to ensure that the sequencing run had produced sequences of a satisfactory quality before any other processing was carried out. The initial assessment of the data was used to inform subsequent processing and quality control. Prinseq provides a graphical visualisation of several parameters and the program parameters can be set specifically filter and trim reads on the most suitable combination. Table 5.1 summarises the descriptive statistics initially determined by analysis with Prinseq for the read data.

Table 5.1 Descriptive statistics determined by analysis with Prinseq

Parameter	File 1 (forward reads)	File2 (reverse reads)
Number of sequences	7,081,660	7,081,660
Mean GC content	51.32 ± 5.86 %	51.09 ± 6.50 %
Ambiguity (bases read as N)	0.74%	0.69%
PolyA tails present	1.52%	1.77%
Artificial duplicates	71.85%	65.69%

There were a large number of reads determined to be artificial duplicates. Inspection of the ten highest occurring duplicates revealed them to be 23S and 16S ribosomal sequences, suggesting rRNA removal had been far from comprehensive. Some reads contained ambiguous bases (called as N in the read data) but a large proportion of reads containing ambiguity had only a single N read; 44.9% of all ambiguous sequences in the forward file and 20.8% in the reverse file. A small proportion of reads appeared to have polyA tails as part of their sequence, with the presence of a polyA tail being arbitrarily described as the presence of 5 or more A residues at either the 3’ or 5’ end of a read. Inspection of the size distribution of polyA tails revealed that the majority of these sequences were less than 40bp in

length, with a small number of sequences consisting entirely of A residues, suggesting that efforts to remove polyA tails enzymatically had been effective.

The quality of the forward and reverse files was broadly similar in terms of the criteria set out in Table 5.1. The quality scores for both files exhibited a drop in quality towards the end of the reads which is a normal artefact of sequencing (Loman *et al*. 2012). The reverse file has a much greater drop in quality with the last ten reads of many sequences having quality scores below 20 (Fig 5.1). On the basis of the quality information revealed by this analysis, it was decided to filter the data as follows:

- Remove artificial duplicate sequences
- Remove sequences with an average quality score < 20
- Remove sequences with more than 1 ambiguous base (N read)
- Remove sequences of low complexity, defined as an entropy score below 60
- Trim polyA regions of > 5 nt at the 5’ or 3’ end of a read
- Trim bases with quality scores < 20 at the 5’ or 3 ‘ end of a read

These reasonably stringent filtering conditions were chosen to ensure high quality data would be passed onto downstream applications.

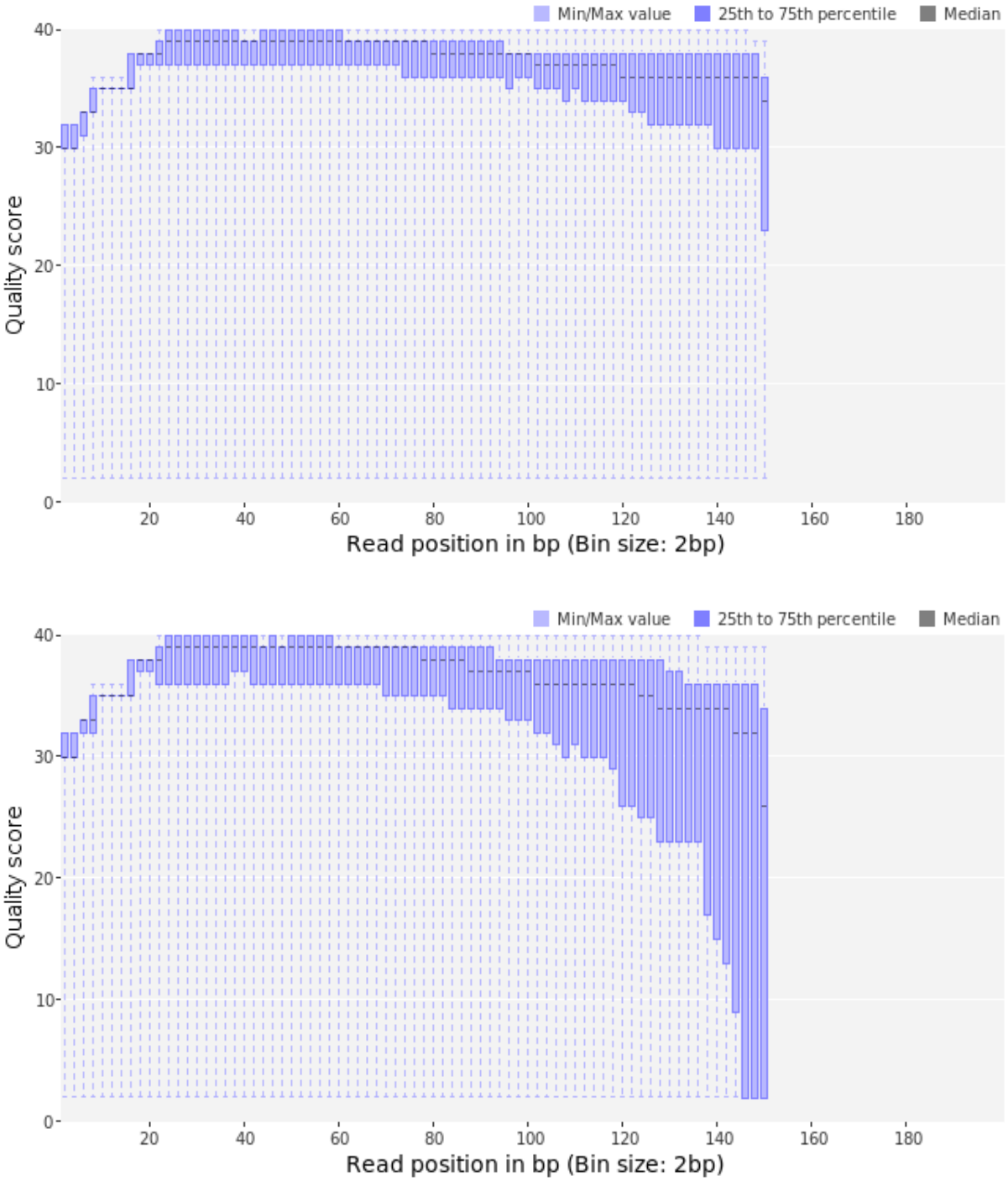


Fig 5.1 The per-base quality scores for the forward and reverse sequence files. Upper graph: Forward reads file quality scores; lower graph: Reverse reads file quality scores. While the majority of sequences had an overall quality score > 20, the last 10 read positions of the reverse reads.

5.5 Assembly with Velvet de-novo assembler and other tools.

It was decided that assembly was an option worth exploring for the metatranscriptome. Although the dataset was likely to have low coverage, assembly of larger contigs was nevertheless a possibility and would be beneficial for downstream analysis, providing longer reads for gene prediction. Assembly was attempted using metavelvet (Namiki *et al*., 2012) and oases (Schulz *et al*., 2012), which are both extensions of the velvet assembly software (Zerbino & Birney, 2008). As metavelvet was developed specifically for use with metagenomic data and oases was developed particularly for transcriptomics, these two velvet extensions seemed the best candidates for achieving an assembly of reasonable quality. Even with assembly, it was not expected that many long contigs would be produced from this dataset as a large number of sequences were likely to represent a single instance of a particular fragment of a gene and coverage of large parts of the metatranscriptome would be so fractional as to render assembly impossible.

Assembly was performed using the quality filtered paired-end sequence data produced according to the criteria set out in section 5.4 above. Velvet expects paired end data in a single file rather than split into two separate files for forward and reverse reads. The perl script shufflesequences_fastq.pl which comes bundled with the Velvet program was used to generate a file where paired sequences were in sequential order. Additionally, quality filtering had in some cases removed one sequence of a pair from either the forward or reverse file, leaving a singleton in the other file. This too is unsupported by velvet and singletons were removed using the bundled script select_paired.pl.

The main adjusted parameter was the hash length of the initial velvet process. Generally choice of hash length will greatly affect the quality of the assembly. It is also impossible to know what the best hash length for any given assembly is and an element of trial and error is often required.

The performance of metavelvet and velvet/oases were assessed using several parameters: Number of contigs produced by the assembly, the size of the largest contig, the mean contig size and the N50 metric. The N50 metric is a measure of the quality of an assembly, defined in Miller *et al*. (2010) as “the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly”. The output for the assemblies obtained are summarised in Table 5.2.

Table 5.2 Assembly statistics for different assembly methods.

Assembly Method	Number of contigs	Largest contig size	Mean contig size	N50 value
Metavelvet – default settings	645	1183	241.92	264
Metavelvet – hash length 13	14669	197	38.05	38
Metavelvet – hash length 19	9995	1595	60.41	60
Oases – hash length 19	754	5404	225.61	246

Assembly initially produced a very small number of contigs although their length exceeded that of the individual sequences. Adjusting the hash length and re-running metavelvet increased the number of contigs. A hash length of 13 successfully increased the number of contigs but at the expense of contig size. The mean size dropped from 241.92 to 38.05 bp. A hash length of 19 caused an increase in contig length compared to default settings with a smaller drop in mean sequence size and an increase in the size of the largest contig. The same settings applied to the Oases extension produced the longest contig maximum although Oases only produced 754 contigs in total.

Assembly parameters were not investigated exhaustively here and there was scope for further experimentation but the output from several assembly runs suggested that in the case of this particular dataset there would not be much benefit from attempting to assemble the metatranscriptome. Only very small numbers of contigs or contigs of extremely short length seemed to be attainable here. There were instances of contigs assembled to longer than 1000 bp in length but these assemblies still produced a low average contig length. It was decided that the unevenness of the metatranscriptome was probably a barrier to useful assembly for these data and that the contigs produced probably represented only a tiny fraction of the genetic diversity of the dataset where coverage was highest. Even with an optimal assembly there would probably be a great deal of genetic information left in short reads and analysis of the assembled contigs would never provide a comprehensive overview of gene expression in the data. Therefore, alternative methods for processing the data were sought.

5.6 Merging paired-end sequences to form single longer reads using FLASH.

Tools have been developed specifically for joining together paired-end sequences, to form a single, contiguous, longer read than the individual shorter pairs.

Joining paired end reads is effectively a poor man's assembly, and the generation of longer reads will facilitate downstream analysis in the same way as assembly of reads into contigs. In this case, with 2x150 bp paired-end reads, the maximum length of sequence represented by the two reads is 300 bp, although pairing reads requires some bases of overlap so any two paired end reads of 150 bp each could conceivably be combined to form a sequence of 280-290 bp in a best case scenario. This method will result in many reads not being combined and remaining as single reads as in some cases the paired ends will not overlap if a fragment of the library was longer than 300 bp, and it was not possible to sequence in from both ends and end up with an overlap. Additionally, short overlaps of 2-3 bp cannot be used to pair sequences reliably either, as such a short overlapping section could easily be coincidental. In spite of some disadvantages, pairing of paired-end reads is a fast, easy process which is much less technically and computationally demanding than full assembly. Given the small number, and short length, of contigs produced by assembly, pairing of reads seemed to be a better way to process this dataset.

The FLASH (**F**ast **L**ength **A**djustment of **SH**ort reads) program was chosen to perform the read-merging step. This software has been described by Magoc *et al*. (2011) and is proven to be fast and accurate; it was therefore considered a reliable tool for generating longer sequences from the paired-end reads. As FLASH merges sequences in fastq format with quality scores encoded, and outputs in the same format, it was decided to run the program on the raw reads, without quality filtering, and then perform filtering on the output. This also avoids the problem of FLASH not supporting missing sequence pairs. FLASH was executed with the following parameters set to determine the merging of sequences:

- Minimum overlap length of 10 (default)
- Mismatch ratio of 0.3. (This is the maximum allowed ratio of the number of mismatches and the overlap length).
- Average read length 150.
- Average fragment length 300.
- Standard deviation of fragment length 30 (set to 10% of the average fragment length as suggested in FLASH documentation).

FLASH output was ca. 3.2 million merged pairs and ca. 7.8 million sequences which could not be merged; 45% of the paired end sequences were combined into a single sequence. The paired and unpaired sequences resulting from running FLASH were quality-filtered using the Prinseq program as detailed previously (section 5.4).

The length distribution obtained by merging paired reads with FLASH is presented in Fig. 5.2.

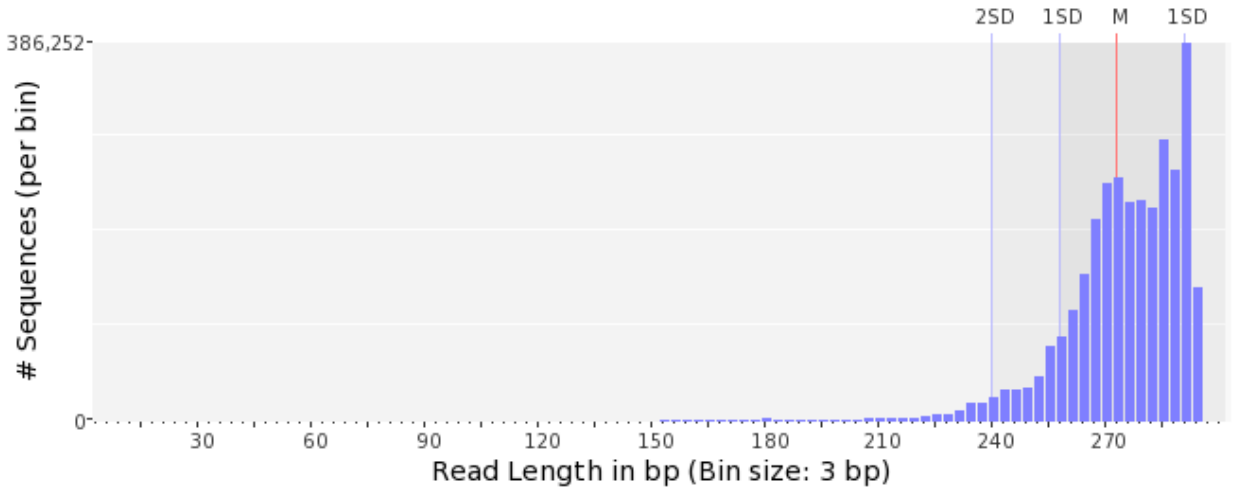


Fig 5.2 Length distribution of the merged paired end reads after FLASH processing. Mean sequence length was 272.83 bp ± 17.4. The modal length was 291 bp. A few sequences overlapped almost entirely.

5.7 Removal of rRNA sequences.

After pairing and quality control the other necessary processing step which was required before functional analysis of the data was the identification and removal of rRNA sequences. Even though the sequencing was performed on a sample that had been rRNA subtracted, this step had significantly lowered but not eliminated the rRNA sequences from the datatset. Previous metatranscriptomic studies have demonstrated that even with rRNA removal and a depletion step implemented into the library preparation pipeline, a large number of the sequenced reads can be rRNA sequences, with proportions from different studies ranging between 50 and 80% (Stewart *et al*., 2010). Ribosomal sequences should be omitted as far as possible from analysis pipelines that screen for protein coding regions as they may result in misannotations and yield spurious gene calling and functional prediction (Tripp *et al*., 2011). Care was therefore taken to ensure that data were thoroughly screened for ribosomal sequences and the vast majority removed.

The tool chosen for rRNA removal here was Ribopicker (Schmeider *et al*. 2012) with the analysis carried out via the webserver. The reads were compared to the Silva small subunit (16S/18S) and large subunit (23S/28S) reference databases (Pruesse *et al*. 2007), which are a high-quality, curated resources. Criteria for classification of reads as rRNA were:

- 75% identity to a sequence in the reference database.
- 40% alignment coverage (i.e. the fraction of the sequence which aligns to a sequence in the reference database).
- Minimum alignment length of 30 bp.

Reads which satisfied these criteria were removed from the data as likely rRNA sequences. These settings were considered to be relatively stringent. Ca. 3.7 million sequences were submitted to Ribopicker and 71.7% of these were found to be rRNA-like and were removed from subsequent analyses.

5.8 Quality Control and data pre-processing: concluding remarks.

After quality filtering and rRNA removal, one final quality control step was carried out. Removal of low quality bases and polyA tails had left some short sequences in the dataset and remaining reads that had passed other criteria but were < 100 bp in length were removed. Shorter sequences are less likely to generate meaningful alignments for annotations. Pairing single reads together to form a single longer sequence, removal of reads of a ribosomal origin and reads of low quality or excessively short length ultimately reduced 14 million individual reads in the raw

dataset to 978,043 quality filtered sequences. This produced a much more manageable, and rigorously filtered, dataset.

5.9 Comparison to other metatranscriptomic datasets.

Comparatively few metatranscriptomic projects have been carried out utilising the Illumina platforms. A few more have been performed using 454 pyrosequencing but published metatranscriptome datasets remain a distinct minority; environmental projects of this type usually target metagenomic DNA or amplicon libraries. The 454 platform, which can produce longer reads than Illumina chemistry of an average length of 800 bp, does generate a much greater range of sequence lengths. The average read length reported from metatranscriptomic sequencing on the 454 platform is not much greater than the maximum achievable by 150x150 paired-end sequencing on the Miseq platform. Mean length of merged paired end reads was 272.83 bp and mean length of the the combined high-quality merged and unmerged reads was 195.21 bp. Pyrosequencing of cDNA libraries of metatranscriptomes has yielded read lengths of 99-97 bp (Shi *et al*. 2011) and 161-208 bp elsewhere (Stewart *et al*., 2012). Therefore, Illumina paired end sequencing can, by making use of freely available software, produce read lengths at or exceeding the read lengths produced by pyrosequencing. Given the numbers of non-rRNA reads reported from these studies (69,200 – 268,093 by Stewart *et al*. (2012) and 40,760 – 68,712 by Shi *et al*. 2011) the use of Illumina sequencing in this project has increased the amount of potentially useful data tenfold.

The volume of rRNA reads in the dataset (71%) accounted for a large chunk of the raw data but for metatranscriptomic sequencing projects as a whole, this level of rRNA is not at all unexpected. Stewart *et al*. (2010) reported an effective ribosomal subtraction method which they validated by comparing the results of pyrosequencing metatranscriptomes from seawater that had been rRNA-subtracted and rRNA-unsubtracted. 80-88% of the unsubtracted samples consisted of rRNA-like reads, while subtracted samples were found to have an rRNA content of 52-61%. This was achieved using a tailored hybridisation method, with custom made rRNA probes as described in chapter 4. The enzymatic rRNA treatment used to produce the sample in this study was a much less labour intensive, and less expensive procedure and a value of 70.1% rRNA is a significant improvement on the amounts of unwanted rRNA reported for unsubtracted metatranscriptomes. In another study, rRNA reads comprised 37.1% – 58.1% of the total (Shi *et al*. 2011). No specific rRNA subtraction method was employed in this case so the proportion of rRNA reads

is perhaps surprisingly low. This result also suggests that rRNA content in a metatranscriptome might be related to factors such as the extraction process itself, the nature of the environmental sample and the metabolic state of the microbial population in that environment.

5.10 Data mining

In order to identify gene-coding sequences in the dataset and investigate the functional sequences in the metatranscriptome, several analytical pipelines were used. An overall analysis was obtained using the MG-RAST webserver and the MEGAN software package. Both of these provide an excellent view of the community in terms of a general summary of the functional genes present in a metagenome or a metatranscriptome, as well as a breakdown of the taxonomy of the dataset. These tools work in rather different ways. MEGAN requires a blast output file which it uses to generate a visual summary of the taxonomic classification of the reads, and where possible sequences are mapped to functions identify sequences which correspond to SEED subsystems and KEGG orthologies. The MG-RAST webserver uses an analysis pipeline to predict protein-coding ORFS and rRNA features. Summaries of the data are produced for taxonomy and for predicted proteins that could be annotated and assigned to a functional category.

5.10.1 Blastx search run to provide an output for MEGAN analysis.

A Blastx search of the 978,043 quality-filtered sequences was initiated against the NCBI non-redundant (nr) protein database using an E-value cutoff of -10^{-3}. Given the large number of sequences, in order to increase the speed of the blast run and simplify the output file, the search was limited to one result per query sequence.

The output from this blast search was imported into MEGAN for analysis. A summary of the blast output as summarised by MEGAN classification is presented in Fig 5.3. Using the nr database as a target for the blastx search is slow, due to the sheer size of the database. The main advantage in using a relatively slow blast search against a database of this size is that the sensitivity of the algorithm combined with the comprehensiveness of the nr database represents an effective way of assigning an identity to a gene from an environmental sample where a large number of the species present are likely to be poorly represented in databases generally. Smaller, high-quality databases (e.g. Swissprot ref) exist which are faster to search against can also be used. These are manually curated, but as manual curation is based on experimental data and careful inspection of sequence and

prediction information, such databases are likely to be limited to more well-studied, better-characterised organisms. Using a larger database which includes many sequences whose characteristics are based on automatic predictions, or which are “hypothetical” proteins determined from sequencing projects, might make for more imprecise assignment of function based on homology but is more likely to produce informative hits when working with sequences of an environmental origin where the organisms represented by the dataset might have few, if any, cultured reference strains.

Even searching against the nr database, a large number of sequences were not found to have a match at the E-value cut-off used here. The largest top-level category of blast hits assigned by MEGAN was “No Matches To Database” which represents 40.4% of the quality-filtered metatranscriptome (Fig. 5.3). The next largest, Domain Bacteria, constituted 35.1%. Although it is possible that many of these sequences could have been matched to the database with a relaxation of the stringency of the search, this would also increase the proportion of incorrect hits resulting from query sequences having similarity to sequences in the database by chance alone and not due to any ancestry. These results represent a trade-off between accuracy and comprehensive coverage.

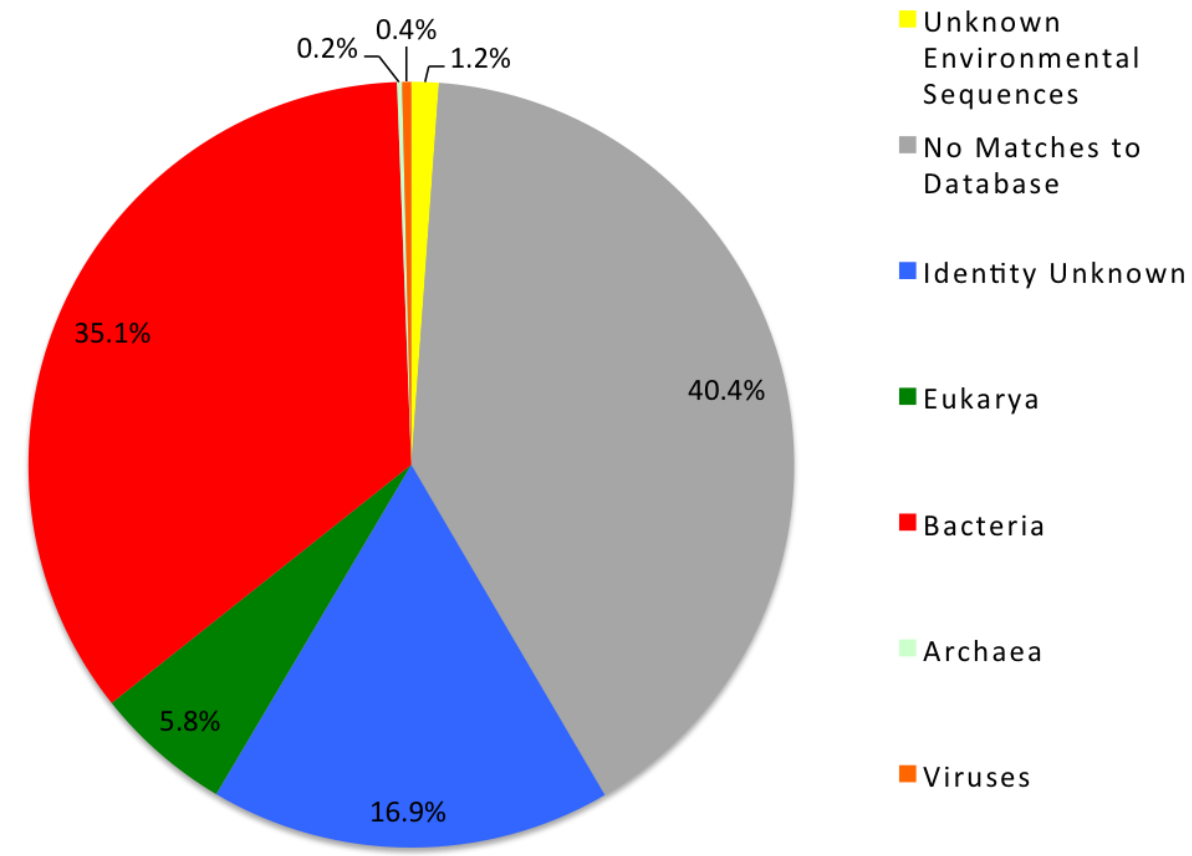


Fig 5.3 Classification of blastx results by MEGAN at the program's top level; Domains, or categories to which sequences are assigned if a Domain cannot be determined.

5.10.2 MG-RAST and MEGAN analysis of the metatranscriptome.

Functional and taxonomic classifications were carried out with MEGAN and MG-RAST. The two analysis pipelines produced similar output in some respects, with the taxonomic assignments particularly well matched. In other respects, the output from MEGAN and MG-RAST analysis yielded different results, especially in terms of numbers of sequences assigned to functional categories.

An overview of the taxonomic assignment by these two methods is presented in Fig. 5.4a. The total numbers of sequences assigned to the domains Bacteria, Archaea, Eukarya, Viruses and to a miscellaneous category ("Other") for reads assigned as similar to environmental sequences of unknown origin in the database are compared. Reads in the "Other" category could be from any of the other domains, but their identity cannot yet be determined. The general trends are preserved by both classifications. The Bacteria is the dominant domain in the metatranscriptome, followed by the Eukarya. The Viruses are the next most abundant domain; MG-RAST classified many more reads as viral in origin than MEGAN (21735 compared to 3790) and MEGAN classification recorded more unknown environmental sequences than viral reads, although some of the reads determined as unknown by MEGAN may have been classified as viral by MG-RAST. There were in fact more reads classified as unknown environmental origin than viral by MEGAN but this still leaves Viruses as the third largest true domain. The Archaea were in an extreme minority according to MEGAN, being represented by only 2037 reads. MG-RAST detected no sequences with significant matches to Archaea at all.

The abundance of each domain as a percentage of the total number of reads that could be classified to domain level or as "Unknown Environmental" is displayed in Fig. 5.4b. That Bacteria contribute the greatest part of the metatranscriptome and that the Eukarya are a clear second is in no doubt although classification of viral reads highlights a discrepancy where MG-RAST has identified 7.38% of its total assigned reads to the Viruses Domain compared to 0.91% of total reads assigned by MEGAN. MG-RAST therefore indicates a much greater presence for viruses in the community. Overall the MG-RAST and MEGAN analysis produced similar results for the taxonomic classification at Domain level. MEGAN was able to assign more reads overall to either a Domain or as miscellaneous sequences for which a Domain-level classification could not be determined (416708 reads assigned by MEGAN compared to 294687 assigned by MG-RAST).

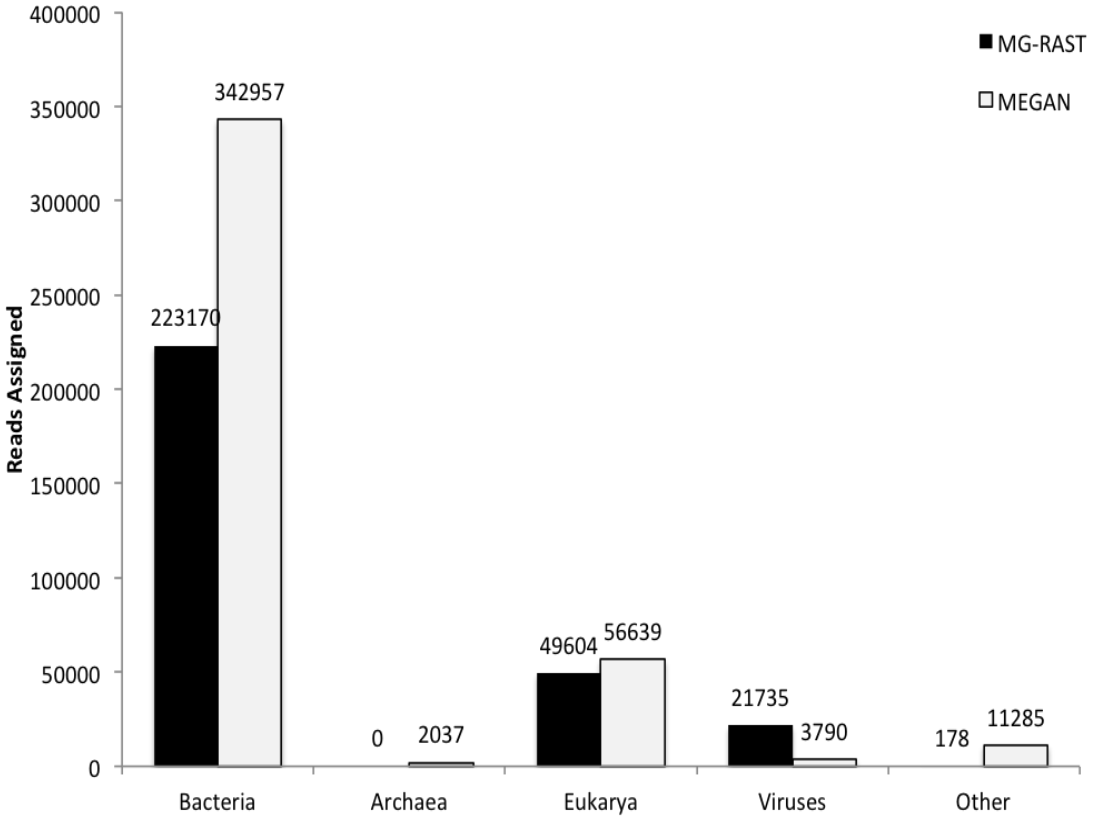


Fig 5.4a Numbers of reads assigned at Domain level by MG-RAST and MEGAN

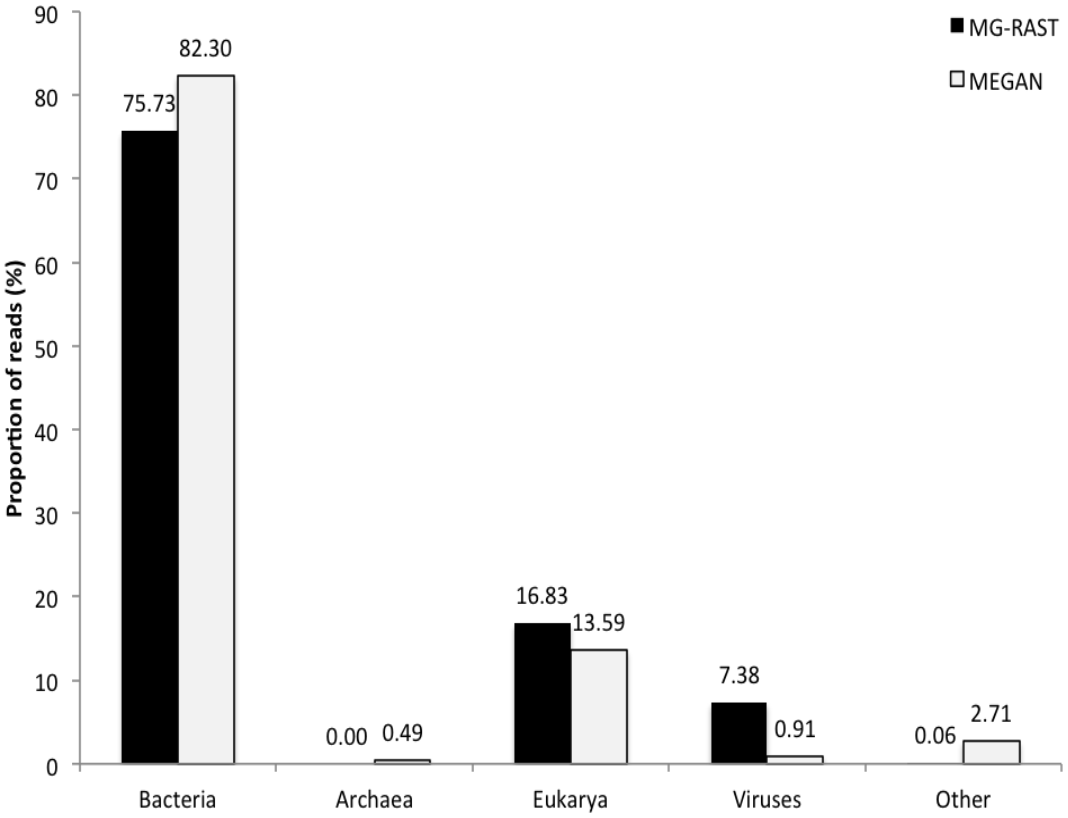


Fig 5.4b Proportion of total reads assigned at domain level by MG-RAST and MEGAN

5.10.3 Phylum level diversity of the Bacteria sequences in the metatranscriptome

The Bacteria were significantly more abundant than the other domains, and at Phylum level it can be seen that certain phyla were far more abundant than others i.e. the diversity is very uneven. The abundance of each Phylum as a percentage of the total number of Bacteria determined by both MG-RAST and MEGAN is presented in Figure 5.5. As with Domain level classification, there are differences in the numbers produced by the different analysis pathways but the overall pattern is consistent. The Bacteroidetes, Firmicutes, Proteobacteria and Actinobacteria were the most numerous phyla by a large margin. Numbers of reads assigned to these four phyla were similar, except in the case of the Proteobacteria where a notable discrepancy exists, with 38.19 and 22.03% of the reads assigned by MEGAN and MG-RAST respectively. In addition to these four well-represented groups, there were a large number of phyla comprising much smaller proportions of the community reverse transcribed transcriptome (< 5%, and and often <1%).

The phyla Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria are large groupings of organisms with many representative species and are often found to be highly numerous in metagenomic and metatranscriptomic analyses (Xiong *et al*., 2012; Ulrich *et al*., 2008). Members of these phyla are more often than not found to be predominant in microbial communities.

A total of 30 phyla were detected by the two classifications and 29 of these phyla were common to both. Reads were assigned to the Phylum Candidatus Poribacter only by MG-RAST, and comprised 0.005% of total Bacterial reads; assignments to Caldiserica were made only by MEGAN and constituted 0.003% of the total. A handful of other phyla had assignments from both methods at very low levels, which may indicate that these sequences were picked up as incidental organisms that were not particularly active members of the community.

Given that this metatranscriptome is derived from a community colonising a crystalline cellulosic substance, it follows that Firmicutes should be present in large numbers. This Phylum contains many specialized cellulose degraders, such as members of the genera *Clostridium* and *Ruminoccus* (Lynd *et al*., 2002). The Phylum Bacteroidetes too contains lineages of bacteria renowned for their ability to break down cellulosic material, for example species of the genus Cytophaga (Lynd *et al*., 2002). The Phylum Fibrobacteres, containing the single genus *Fibrobacter*, is also a lineage of cellulose degrading bacteria. Both MG-RAST and MEGAN classifications determined a presence of Fibrobacteres but according to MG-RAST the Phylum is a significantly more important part of the community (1.59% of the total Bacterial population) than revealed by analysis with MEGAN (0.12%). The MG-RAST

assignment suggests that members of the Fibrobacteres might be a small but active part of the community here, a result which mimics the findings of chapter 3 in terms of evidence. Additionally, given the diversity represented by the other phyla and the previously established cellulose degradation exhibited by fibrobacters, the detection of these sequences could reveal a colonization of cellulose-rich cotton by cellulose degrading environmental fibrobacters with much of the diversity of the other highly represented phyla perhaps resulting from members of the biofilm community that are not involved in the primary degradation of the cellulose.

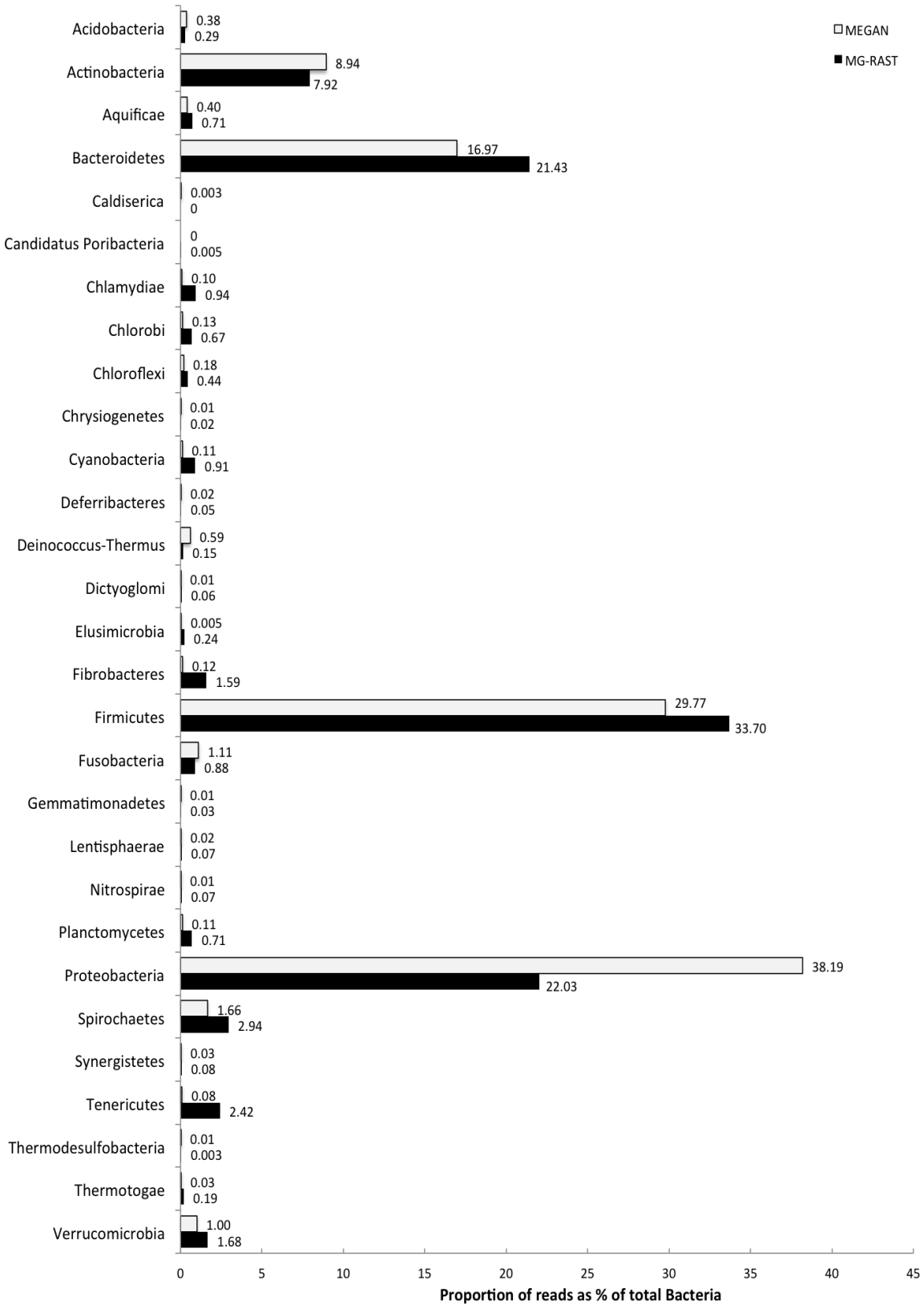


Fig 5.5 Phylum-level breakdown of metatranscriptome sequences.

5.10.4 Efficient high-speed phylogentic classification of read with Metaphlan.

Metaphlan was originally developed as high speed sequence classifier, intended to deal with very large metagenomic datasets (Segata *et al*., 2012). It compares reads against a database of reference sequences that represent markers capable of unambiguously identifying closely matching sequences as being of a particular clade. Metaphlan has been employed for dealing with metagenomic studies of the human microbiome (Huttenhower *et al*., 2012). Its reference sequence database on which classifications might then be somewhat biased towards well-studied organisms with sequenced genomes and be less suitable for the classification of reads from other environments.

The performance of Metaphlan was assessed using the metatranscriptome dataset. The classification obtained using this program differed greatly from that obtained via MEGAN or MG-RAST. Bacteria were determined to be 96% of the community; 65% of the reads were classified to the phylum Bacteroidetes (compared a classification of 16% and 21% of bacterial reads to the Bacteroidetes by MEGAN and MG-RAST respectively). It seemed likely that a large number of reads in the metatranscriptome had no representative sequence in the Metaphlan reference database, which was causing a bias in the classification. It was decided that this method was not able to yield a reliable classification for this dataset. Using bowtie2 to perform an initial alignment did however render Metaphlan an extremely rapid classifier. Possibly, future updates of the reference sequence will increase the utility of the program for datasets containing poorly represented organisms.

5.11 Functional Community Overview

Both MG-RAST and MEGAN are able to provide a functional overview of metagenomic and metatranscriptomic data. Both analysis pipelines are capable of SEED subsystem and KEGG Orthology (KO) classification. MG-RAST additionally provides a classification of reads by COG categories. MG-RAST assigned many more reads to functional categories than MEGAN was able to.

SEED Subsystem classification of the metatranscriptome reads by MEGAN and MG-RAST are presented in Fig 5.6. MG-RAST annotation assigned 10,229 sequences to the SEED category 'Phages, Prophages, Transposable elements, Plasmids' which suggests that phage or mobile genetic elements might be active in the sediment microbial community. The majority of the reads assigned to this category were classified due to high homology to phage capsid proteins. MEGAN analysis, however, assigned only 6 sequences to this category. The MG-RAST annotation has sequences assigned in relatively high numbers to subsystems related to amino acid biosynthesis and metabolism of proteins, RNA, carbohydrates and lipids. Respiration is also well-represented. Within the categories, some well represented sub-categories include Bacterial small ribosomal subunit protein biosynthesis which constituted 1270, or 13% of the 8357 reads assigned to Protein Metabolism by MG-RAST. Overall, this suggests an active community where protein and mRNA turnover is ongoing concomitant with cells growing and multiplying, and a community subject to significant predation by phage.

Reads assigned to KEGG Orthologies are summarised in Fig. 5.7 and the pattern that emerges here is similar to that seen for the Subsystems categorisation. MG-RAST once again assigned many more sequences than MEGAN, although in both cases the highest number of classifications was to the Metabolism category, which represents activities such as carbohydrate and lipid metabolism, and amino acid biosynthesis. The second highest number of classifications was to the Genetic Information Processing category, which represents Transcription and Translation functions. This is again consistent with an active community of microorganisms, growing and dividing.

MEGAN assigned far fewer sequences to the KEGG Orthologs (KO) and SEED subsystems functional categories. This difference is most likely down to differences in how the data is handled by the MG-RAST pipeline and the MEGAN classification system. MEGAN attempts to perform functional assignments based on the highest-scoring blast hit. Given that the blast search was limited to one hit per query there were probably very few query hits where a definitive functional role could be described, making it impossible for the program to determine functional roles for

most of the metatranscriptome. Increasing the number of blast hits slows the speed on the search, and does not absolutely guarantee a hit to a protein with a defined functional role. MG-RAST uses gene prediction, and compares sequences directly to a database containing KEGG and SEED annotations.

These functional summaries provide only the broadest of overviews of the gene expression of the microbial community. There are a large number of subcategories, and many functions grouped within each subcategory. Many of the functions described by KEGG and SEED are housekeeping functions, which will be found to be “always on” across the Bacteria and are in no way specific to a cellulose degrading community.

It is also necessary to take functional classifications based on homology to genes with a defined function as a best guess result, and not as certain confirmation of the presence of a particular function in the metatranscriptome. Consider the classification of reads to the KO Human Diseases. Assignments to this KO by MG-RAST constituted 1% of total KEGG assignments, which is a small proportion. However, the presence of genes involved in human virulence at the bottom of a lake is somewhat surprising. 47% of these were classified as being related to infectious disease in humans which is somewhat more understandable as bacterial species in the environment often harbour virulence factors that can allow them to be opportunistic pathogens of humans under the right circumstances. Additionally, genes involved in secretion system pathways or iron acquisition might be classified as “virulence factors” even if they are of utility in the wider environment generally and helpful when competing other bacteria and not so much the human immune system. Classifications of sequences from a bacteria-dominated lake sediment metatranscriptome as being related to human cancers and neurodegenerative diseases is more puzzling, but the explanation probably lies in that certain enzymes with specific E.C. numbers are implicated in the development and progression of these diseases and sequences from unrelated organisms with similar activities in their respective species may end up being classified in these groups. With this in mind it is better to regard functional categorisation of this type as a useful guide, able to reveal interesting general characteristics that would benefit from deeper analysis.

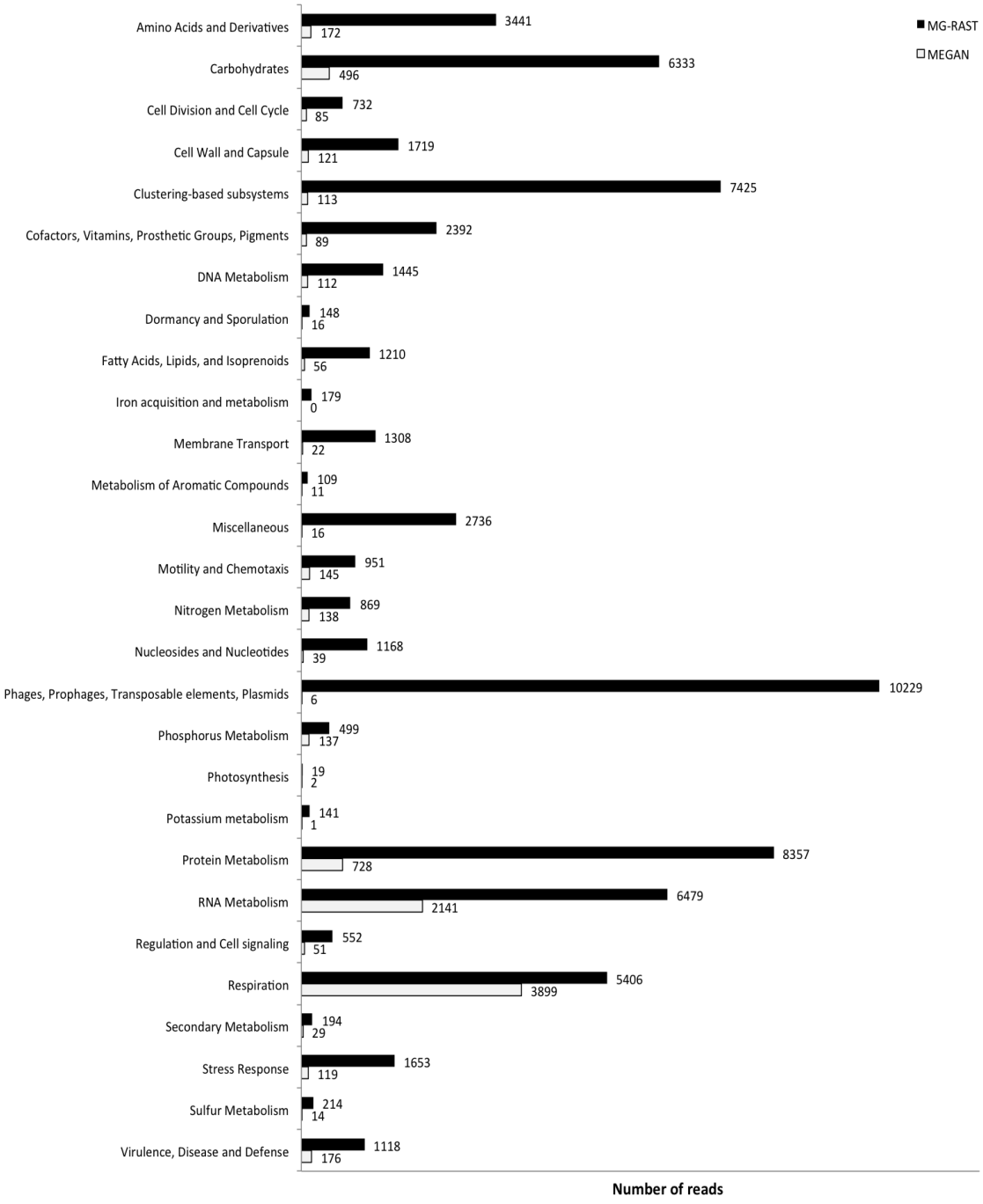


Fig. 5.6 Classification of metatranscriptome reads to SEED categories

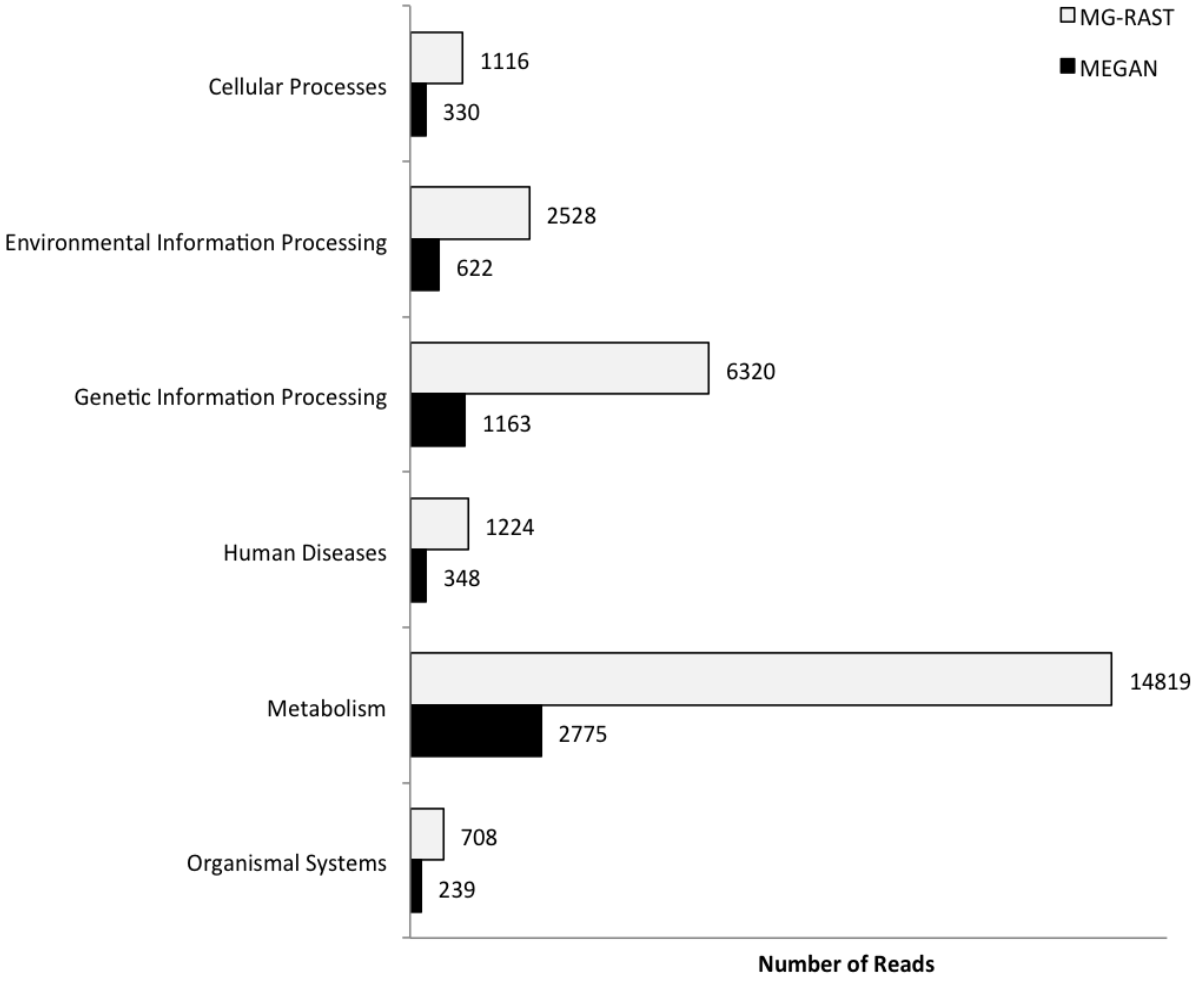


Fig 5.7 Classification of metatranscriptome reads to KEGG orthologies.

5.12 Genome level functional assignment by MG-RAST analysis.

The MG-RAST analysis tools allow matches to particular genomes to be viewed. Recruitment plots give an indication of how thoroughly the entire genome is covered by the metagenome or metatranscriptome and an indication of the levels of coverage of individual features within that genome.

The genome to which more reads had been mapped than any other was in fact a phage genome, specifically Enterobacteria Phage PhiX174 *sensu lato*, to which 10,170 sequences has been ascribed. These sequences provided coverage of almost the entire genome, forming two contigs and therefore very nearly providing a full assembly of the genome. The coverage was uneven, which probably reflects the expression profile of the phage genes. The most hits corresponded to the capsid protein which is required in greater numbers than the others as sixty copies are needed in the assembly of individual phage particles (Roux *et al*. 2012). Details of the mapping to the genome of this phage are presented in Fig 5.8 and Table 5.3. The average identity of the hits to the sequence they were mapped to was quite high, with the lowest at 95.8% identity and the highest at 99.04%. At this level of identity it seems that the mapping is probably correct but some sequences could originate from different, but related, phages with high levels of conservation in their genome and protein structures. The only protein not represented in the metatranscriptome is the lysis protein.

Taxonomic classification had suggested that fibrobacters might be present in the community and MG-RAST has additionally mapped reads from the metatranscriptome to the genome of *F. Succinogenes* S85. Summaries of the mapped reads are presented in Fig. 5.9 and Table 5.4. There are many hits to housekeeping genes involved in overall cell metabolic functions; the largest number of hits corresponded to an RNA polymerase subunit. Detection of genes involved in transcription and translation suggests that fibrobacters might be relatively active in the community and reads were also mapped to glycoside hydrolase family member proteins indicating potentially active degradation pathways being expressed to attack the cotton cellulose. Identity scores of mapped reads were lower here than for the phage genome mapping which might indicate the presence of relatives of *F. succinogenes*, environmental cousins of an organism though to exist exclusively in the rumen. It should be noted that the only sequenced member of the fibrobacteres is *F. succinogenes* S85, and sequences from related, uncultivated environmental species from this group are likely to be matched to this genome, having no better representative sequence.

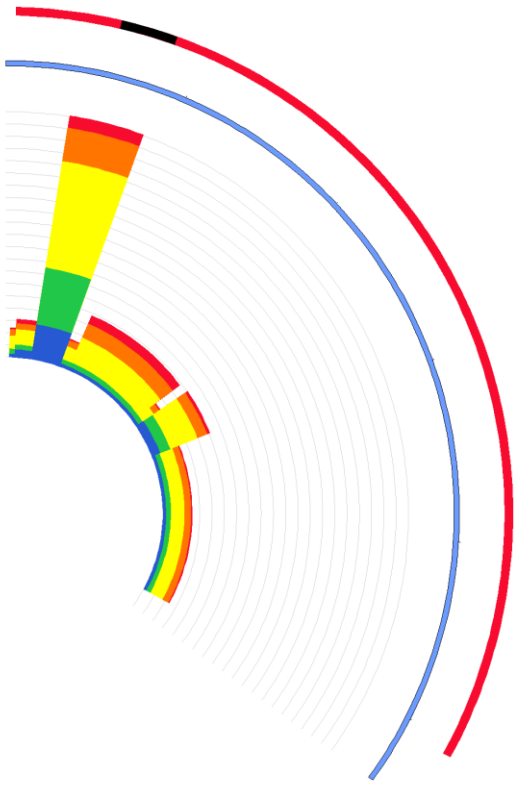


Fig 5.8 Visual summary showing the mapping of reads to the Enterobacteria Phage PhiX174 *sensu lato* genome by MG-RAST. The light blue band represents the length of the genome. The outer red and black band represents annotated protein coding features; red signifies that reads were matched to that feature, black represents a feature with no reads mapped to it (specifically the lysis protein E, the only protein encoded in the phage genome not represented in the metatranscriptome). The inner stacked bar represent the numbers of sequences matched to individual annotated features, with colour indicating E-value score; blue represents a score of 10^{-3} – 10^{-5}, green 10^{-5} - 10^{-10}, yellow 10^{-10} – 10^{-20}, orange 10^{-20} – 10^{-30}, red 10^{-30} and less.

No. of hits	Function	Average Identity of hits (%)	Average alignment length (aa)
2625	capsid protein	98	46.966
1396	DNA replication initiation protein gpA	99	56.506
1212	protein A	98	42.527
1211	major spike protein	96	39.499
1158	external scaffolding protein	98	39.518
1155	minor spike protein	96	49.909
853	internal scaffolding protein	98	41.254
538	C	98	36.645
141	protein K	99	24.451
25	DNA packaging protein	98	22.092

Table 5.3 Summary of reads mapped to the Enterobacteria Phage PhiX174 *sensu lato* genome by MG-RAST

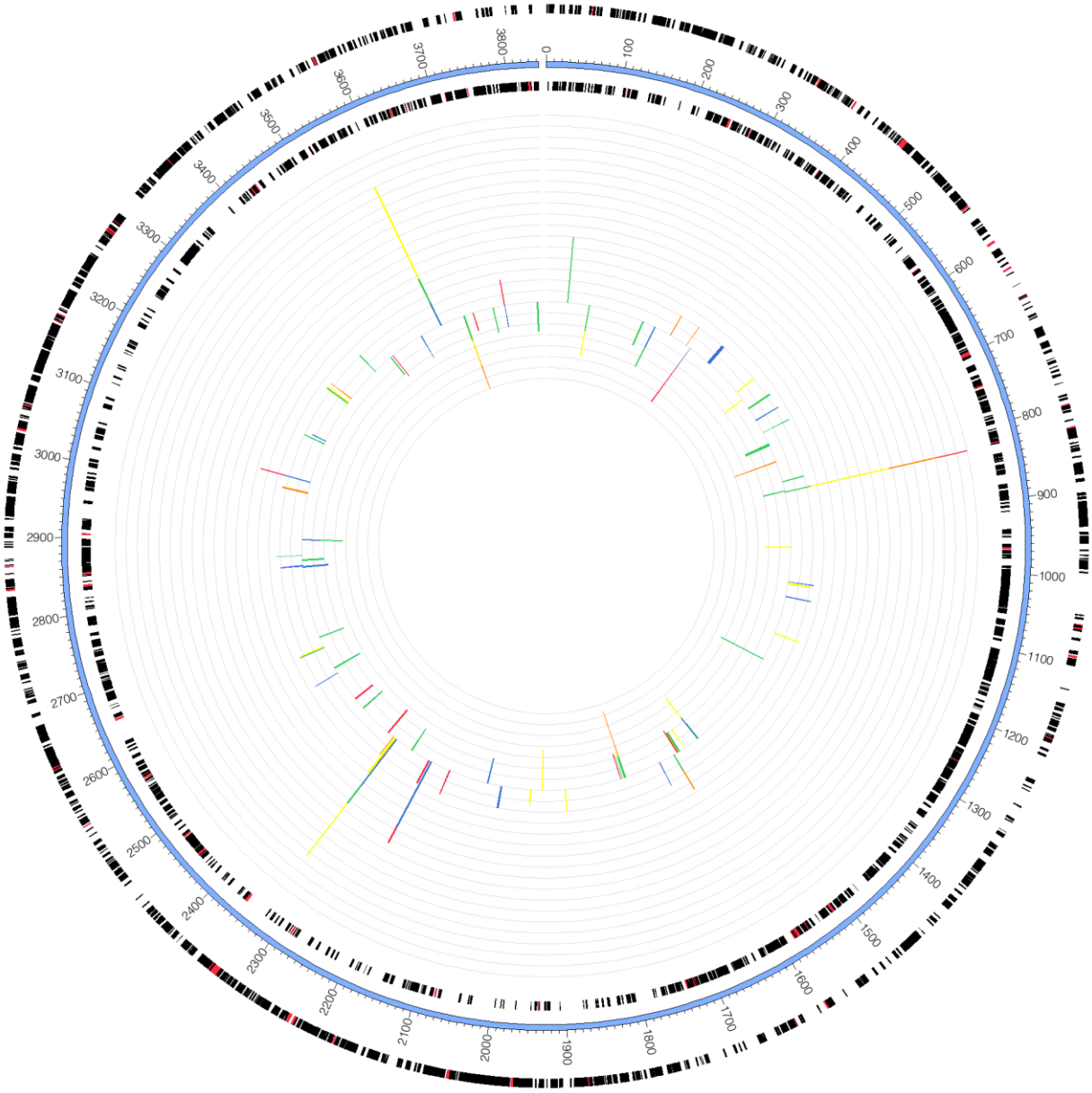


Fig. 5.9 Visual summary of mapped reads to the *Fibrobacter Succinogenes* S85 genome by MG-RAST. Blue ring is genome position in kb, outer and inner black rings are annotated features, positive and negative sense respectively. Stacked bars in the two inner rings represent the numbers of sequences matched to individual annotated features, with colour indicating E-value score; blue represents a score of 10^{-3} – 10^{-5}, green 10^{-5} - 10^{-10}, yellow 10^{-10} – 10^{-20}, orange 10^{-20} – 10^{-30}, red 10^{-30} and less.

No. of hits	Function	Average Identity of hits (%)	Average alignment length (aa)
44	RNA polymerase, sigma 32 subunit, RpoH	78	59
16	cellodextrin-phosphorylase	71	44
7	OmpA/MotB domain protein	74	37
6	glycoside hydrolase family 8	84	16
5	DNA-directed RNA polymerase, alpha subunit	68	71
4	argininosuccinate synthase	77	81
4	glycoside hydrolase family 9	73	30
4	hypothetical protein	83	19
3	D-alanine/D-alanine ligase	69	42
3	glyceraldehyde-3-phosphate dehydrogenase, type I	86	40
3	hypothetical protein	88	17
3	translation elongation factor Tu	91	86
3	S-adenosyl-L-homocysteine hydrolase	77	52
3	preprotein translocase, SecY subunit	71	46
3	translation elongation factor Tu	91	86
3	phosphoenolpyruvate carboxykinase	93	42
2	type II and III secretion system protein	67	34
2	DNA polymerase III, subunits gamma and tau	77	62
2	cysteine synthase A	69	48
2	ABC transporter related protein	61	42
2	extracellular solute-binding protein family 5	69	37
2	hypothetical protein	70	27
2	ribosomal protein S3	62	41
2	glycoside hydrolase family 5	65	56
2	hypothetical protein	63	57
2	TPR repeat-containing protein	78	48
2	FG-GAP repeat protein	76	89
2	hypothetical protein	88	16
2	putative type II restriction enzyme (methylase subunit)	72	74
2	S-adenosylmethionine synthetase	78	91
1	Glycosyl hydrolase family 98 putative carbohydrate binding module	75	36
1	pentapeptide repeat protein	69	78
1	glycoside hydrolase family 9	86	21
1	elongation factor G	90	49
1	DNA topoisomerase IV subunit A	89	27
1	DNA topoisomerase IV subunit B	67	30
1	hypothetical protein	71	35
1	excinuclease ABC, A subunit	55	44
1	pectate lyase	62	42
1	glycosyl transferase group 1	64	39
1	Malate dehydrogenase (oxaloacetate-decarboxylating) (NADP(+)) Phosphate	86	49

	acetyltransferase		
1	RNA polymerase, sigma 32 subunit, RpoH	77	22
1	DNA gyrase, B subunit	74	46
1	glucosamine/fructose-6-phosphate aminotransferase, isomerizing	68	28
1	nucleotide sugar dehydrogenase	80	44
1	binding-protein-dependent transport systems inner membrane component	68	44
1	ribose-phosphate pyrophosphokinase	67	51
1	DNA-directed RNA polymerase, beta subunit	66	41
1	ribosomal protein L1	73	71
1	diaminopimelate decarboxylase	70	46
1	glycoside hydrolase family 5	60	30
1	Mannan endo-1,4-beta-mannosidase	82	17
1	Alpha-galactosidase	88	16
1	isoleucyl-tRNA synthetase	79	87
1	M6 family metalloprotease domain protein	59	37
1	TPR repeat-containing protein	70	90
1	Biopolymer transport protein ExbD/TolR	66	65
1	TonB family protein	74	74
1	hypothetical protein	79	19
1	DNA topoisomerase III	75	32
1	hypothetical protein	83	23
1	coagulation factor 5/8 type domain protein	69	39
1	NADH dehydrogenase (quinone)	75	36
1	NADH dehydrogenase (quinone)	85	39
1	type III restriction protein res subunit	67	46
1	glycoside hydrolase family 10	73	22
1	endo-1,4-beta-glucanase/xyloglucanase, putative, gly74A	66	44
1	exodeoxyribonuclease III Xth	63	41
1	surface antigen variable number repeat protein	60	45
1	ATP-NAD/AcoX kinase	75	32
1	protein of unknown function DUF214	67	39
1	ATPase AAA-2 domain protein	86	49
1	1-aminocyclopropane-1-carboxylate deaminase	70	81
1	GTP cyclohydrolase II	75	32
1	ATP-dependent metalloprotease FtsH	57	44
1	glutamate--cysteine ligase GCS2	87	23
1	hypothetical protein	81	21
1	hypothetical protein	68	34
1	tryptophan synthase subunit beta	86	86
1	ribosomal protein L21	88	25
1	Carbohydrate binding family 11	59	39

Table 5.4 Summary of reads mapped to the genome of *F. Succinogenes* S85 by MG-RAST.

5.13 Rank analysis of Blastx hits.

In addition to MEGAN analysis of a blast search output, a second blastx search against the nr database was conducted to produce a short format blast output file. This output file was used to perform a summary analysis of hits to the blast database by the metatranscriptome to establish which genes in the database were highly represented. For this search, the E-value cutoff was 10^{-3} and ten results per query sequence were allowed. A total of 4,437,626 blast hits were recorded, to a total of 228,203 unique results. The top 10 blast results with the most hits represented 583, 205, or 13%, of the total hits.

The top blast hits are summarised in Table 5.4 below. The annotations for the top 10 matches were, however, not informative as all of these proteins were labeled as "hypothetical" with no function suggested.

Table 5.4 Top 10 blastx hits ranked by the number of query sequences which were matched to each result. Count refers to number of metatranscriptomic reads matched to each

Count	Avg. Identity (%)	Source Genome or Organism	Annotation
97672	53.19	uncultured Rhizobiales bacterium HF4000_32B18	hypothetical protein
86279	63.66	Bacteroides sp. 3_1_23	conserved hypothetical protein
58664	50.59	Streptomyces sp. SPB74	hypothetical protein SSBG_04935
56167	54.42	Streptomyces ghanaensis ATCC 14672	LOW QUALITY PROTEIN: conserved hypothetical protein
54334	62.4	Flavobacteria bacterium MS024-3C	conserved hypothetical protein
49819	63.1	Bacteroides ovatus 3_8_47FAA	hypothetical protein HMPREF1017_03880
48547	59.16	Streptomyces sp. SPB78	LOW QUALITY PROTEIN: conserved hypothetical protein
45310	63.09	Streptomyces viridochromogenes DSM 40736	conserved hypothetical protein
44072	59.77	Streptomyces griseoflavus Tu4000	hypothetical protein SSRG_03841
42341	56.98	Pseudoflavonifractor capillosus ATCC 29799	hypothetical protein BACCAP_04210

5.14 Data mining: Searching for cellulase sequences.

In order to investigate cellulases expressed by the microbial community and present in the metatranscriptome dataset, an analysis pathway was devised based on the method used by Hess *et al*. (2011) to identify glycoside hydrolase sequences from metagenomic data. The quality-filtered metatranscriptome that was submitted to MG-RAST and used for the blastx search for MEGAN classification was uploaded to the Metagenemark webserver (Zhu *et al*. 2010) for open reading frame prediction. Predicted ORFs were used to search against the Pfam database of hidden markov models using the Pfamscan.pl script provided by Pfam (Finn *et al*. 2010). This script uses the HMMER3 software, which is required to be locally installed, to perform a sensitive search against the Pfam HMM database. This search method allows for the identification of predicted ORFs that have a statistically high similarity to Pfam families representing glycoside hydrolase activity and therefore indicating reads in the metatranscriptome that partially encode proteins involved in hydrolysis of polysaccharide molecules and, consequently, with a potential role in cellulose degradation. Classification of a read to a Pfam family informs identity, and possible metabolic role, but does not give an absolute indication of the catalytic function of the encoded protein. For example, an ORF with homology to Pfam family glyco_hydro_3 could conceivably represent a protein with any of the activities related to glycoside hydrolase family 3 which, as listed in the Cazy database, include b-glucosidase, xylan 1,4-b-xylosidase, b-N-acetylhexosaminidase, glucan 1,3-b-glucosidase, glucan 1,4-b-glucosidase, exo-1,3-1,4-glucanase, and a-L-arabinofuranosidase activities.

5.14.1 Pfam database search output.

A fasta file containing the ca. 200,000 putative ORFs predicted from the metatranscriptome by Metagenemark was used as the input file for a search with Pfamscan.pl, against the high-quality curated Pfam-A database. Sequences with hits to glycoside hydrolase (GH) families in the output file were extracted from the fasta file of predicted ORFs for further analysis. A total of 503 sequences were found to have a high level of similarity to a Pfam GH family and to merit further investigation.

Fig. 5.10 summarises hits to GH families as determined by searching against the Pfam database. The majority of GH families detected had only 1-4 hits. The four GH families with the greatest number of hits were GH 3, 8, 9 and the family named Cellulase by Pfam, which in fact represents the GH 5 family according to the description provided in the Pfam database. All GH families contain multiple enzyme activities and assignment of a sequence to a particular family is not an absolute indicator of function. The GH 5 family represents several enzymatic activities

including several specifically related to cellulose breakdown (e.g. endo- and exo-B-1-4-glucanases) and reads with similarity to the representative Pfam family are highly likely to be sequences that encode the active sites of glycoside hydrolase proteins that can attack cellulose, rather than other less recalcitrant polysaccharides.

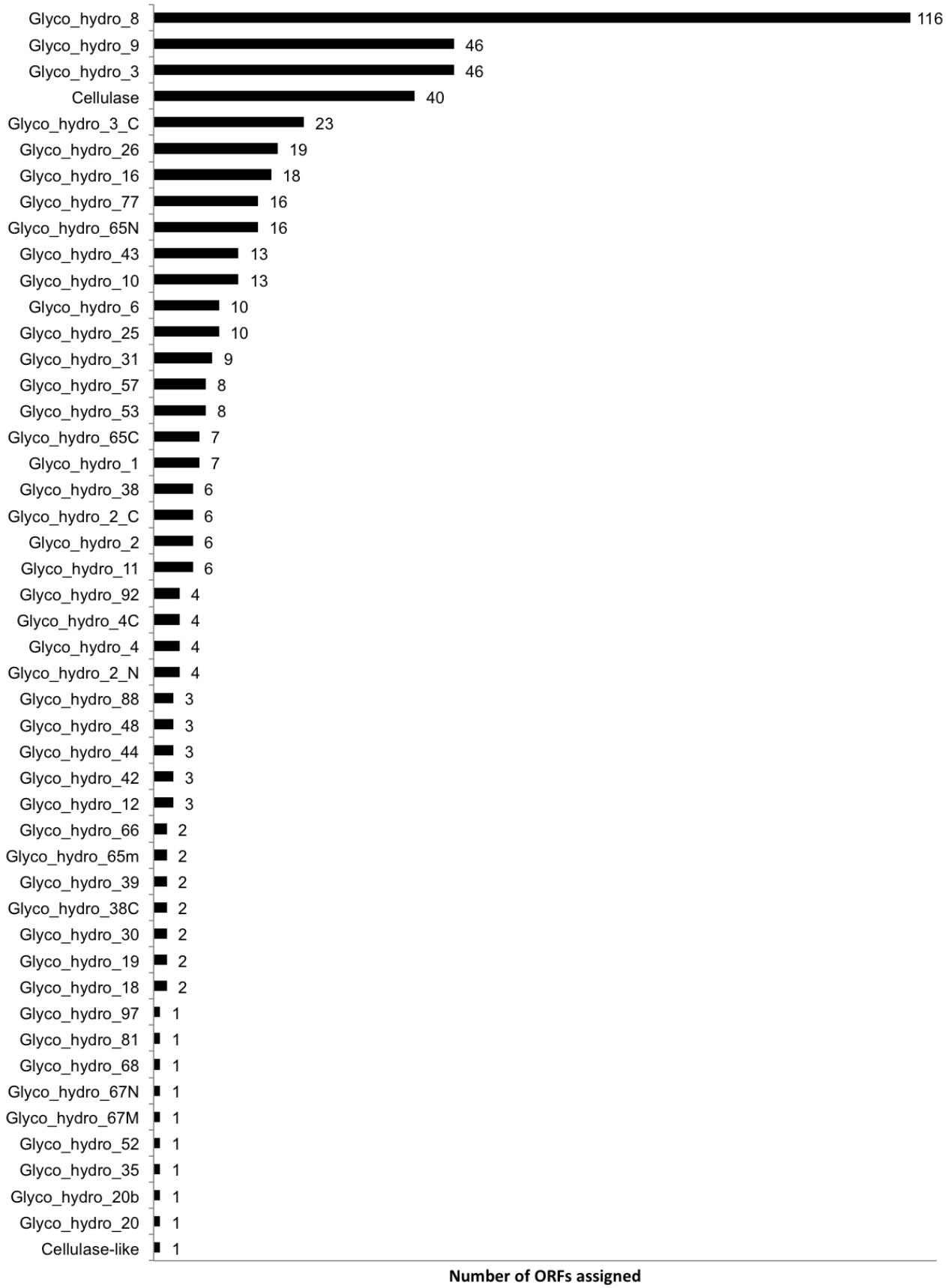


Figure 5.10: Legend follows on next page

Fig 5.11 (Above) Distribution of predicted ORFs across Pfam groups corresponding to glycoside hydrolase families. The Cellulase family was formerly known as "Glyco_hydro_5".

GH families 3, 8 and 9 all contain a range of enzymatic activites many of which can contribute to the breakdown of cellulose, either as part of an initial attack on cellulose polymeric fibres or as downstream enzymes which process breakdown products of cellulose such as cellobiose to glucose monomers. The heterogeneity of the activites of GH families is something of a barrier to interpretation of this data.

The 503 predicted ORFs implicated as representing sequences coding for proteins that were members of Pfam families associated with polysaccharide breakdown were further subjected to blast analysis. A blast search of the 503 sequences was performed against the NCBI NR database as carried out previously for the entire metatranscriptome with one search result per query sequence reported. The results of this blast search are reported in Table 5.5 below. Some of the query sequences corresponded to the same blast hit; 239 of the sequences shared a blast hit with one or more other sequences. There were a total of 473 blast hits reported which means that 30 sequences did not have a hit to the database. These 30 sequences could represent highly novel enzymes for which no relatives could be found on the basis of homology or their relation to Pfam families may have been determined erroneously. The highest number of hits to a single sequence in the database was 20, corresponding to an "endoglucanase-like protein" from the genome of *Cytophaga hutchinsonii* ATCC 33406. Of the top 5 most numerous blast hits, 4 were from this *C. hutchinsonii* genome. Other hits to genes putatively described as polysaccharide hydrolases in this genome were also present in the list.

While the Blast results reveal the closest match in the database for the query sequences, the identities were often low enough that the query sequence might in fact be representative of a related similar, but distinct, function. Many blast hits were to hypothetical proteins and/or proteins originating from uncultured bacteria. Determining a definite function for any of the sequences based on the blast results is therefore very difficult at best. Multiple query sequences with the same blast hit might however indicate that genes expressed in the metatranscriptome of the community colonising the cellulase bait are a specific response to the substrate. The preponderance of relatively low identities of the matches revealed by the blast hits perhaps points to the presence of enzymes not previously charcterised. The extent to which their activities are truly novel or simply variants with little or no functional

novelty cannot be determined. This is essentially a screening exercise to identify candidates for further study.

Table 5.5 (above) Breakdown of the blast results of a blastx search of the 503 ORFs with high homology to Pfam glycoside hydrolase families against the NCBI NR database. Columns represent the number of times a specific entry in the database was a match for an ORF query sequence, the average % identity for the matching ORFs, and annotation information, where available.

No. of Hits	Avg. % Identity	Source Genome or Source Organism	Annotation
20	62	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
14	49	Cytophaga hutchinsonii ATCC 33406	beta-glycosidase-like protein
13	40	Clostridium josui	endo-1,4-beta-glucanase
11	47	Cytophaga hutchinsonii ATCC 33406	beta-glycosidase-like protein
11	63	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
10	47	uncultured organism	putative carbohydrate-active enzyme
9	50	Cytophaga hutchinsonii ATCC 33406	b-glycosidase
8	63	Cytophaga hutchinsonii ATCC 33406	bifunctional acetylxylan esterase/xylanase, CBM4 module, glycoside hydrolase family 10 protein and carbohydrate esterase family 6 protein
5	38	Clostridium papyrosolvens DSM 2782	glycoside hydrolase family 8
5	70	Mucilaginibacter paludis DSM 18603	maltose phosphorylase
5	52	Melioribacter roseus P3M	mannan endo-1,4-beta-mannosidase
4	70	Zunongwangia profunda SM-A87	beta-N-acetylglucosaminidase
4	80	uncultured organism	putative carbohydrate-active enzyme
4	55	Anaerophaga sp. HS1	maltose phosphorylase
4	72	Pelosinus fermentans JBW45	glycoside hydrolase family 9 protein
4	49	Flavobacterium sp. CF136	Por secretion system C-terminal sorting domain-containing protein
3	79	Cytophaga hutchinsonii ATCC 33406	xylanase
3	47	Flavobacteriales bacterium ALC-1	Trehalose/maltose hydrolase (phosphorylase)
3	42	Halothermothrix orenii H 168	beta-N-acetylhexosaminidase
3	71	Dickeya dadantii Ech703	cellulase
3	63	Prevotella ruminicola 23	family 25 glycosyl hydrolase
3	64	uncultured organism	putative carbohydrate-active enzyme

3	63	Mahella australiensis 50-1 BON	cellulase
3	60	Dysgonomonas gadei ATCC BAA-286	hypothetical protein HMPREF9455_01615
3	86	Zobellia galactanivorans	maltose phosphorylase
3	81	Anaerophaga thermohalophila DSM 12881	Beta-glucosidase
3	76	Anaerophaga thermohalophila DSM 12881	glycoside hydrolase family protein
3	46	Pratylenchus vulnus	beta-1,4-endoglucanase, partial
3	44	Alistipes indistinctus YIT 12060	hypothetical protein HMPREF9450_01466
3	56	Acetivibrio cellulolyticus CD2	glycoside hydrolase family protein
2	61	Cytophaga hutchinsonii ATCC 33406	beta-xylosidase/alpha-L-arabinofuranosidase-like protein
2	74	Trichodesmium erythraeum IMS101	Alpha-glucosidase
2	73	Candidatus Solibacter usitatus Ellin6076	glycoside hydrolase family 3 protein
2	69	Flavobacterium johnsoniae UW101	licheninase
2	55	Trichoplax adhaerens	hypothetical protein TRIADDRAFT_21692
2	44	Fibrobacter succinogenes subsp. succinogenes S85	glycoside hydrolase family protein
2	46	Fibrobacter succinogenes subsp. succinogenes S85	glycoside hydrolase family protein
2	56	Haliangium ochraceum DSM 14365	glycoside hydrolase
2	55	Naegleria gruberi	alpha-mannosidase
2	50	Streptomyces pristinaespiralis ATCC 25486	glycosyl hydrolase
2	56	Paenibacillus curdlanolyticus YK9	glycoside hydrolase family 3 domain protein
2	71	Paludibacter propionicigenes WB4	4-alpha-glucanotransferase
2	48	Cellulophaga algicola DSM 14237	cellulase
2	84	Mahella australiensis 50-1 BON	glycoside hydrolase
2	82	Dysgonomonas gadei ATCC BAA-286	hypothetical protein HMPREF9455_00086
2	34	Paenibacillus sp. HGF7	glycosyl hydrolase family 8

2	44	Runella slithyformis DSM 19594	hypothetical protein
2	54	Equus caballus	PREDICTED: LOW QUALITY PROTEIN: lactase-phlorizin hydrolase-like
2	44	Streptomyces sp. W007	glycosyl hydrolase
2	50	Anaerophaga sp. HS1	glycoside hydrolase
2	59	Mucilaginibacter paludis DSM 18603	glycoside hydrolase family protein
2	63	Mucilaginibacter paludis DSM 18603	arabinogalactan endo-1,4-beta-galactosidase
2	63	Clostridium clariflavum DSM 19732	endoglucanase Y
2	45	Ignavibacterium album JCM 16511	beta-glucosidase
2	43	Solitalea canadensis DSM 3403	endoglucanase
2	73	Saprospira grandis DSM 2844	beta-glucosidase-like glycosyl hydrolase
2	48	Caloramator australicus RC3	endo-beta-1,3-glucanase
2	57	Aquimarina agarilytica ZC1	glycoside hydrolase family 8 protein
2	62	alpha proteobacterium IMCC14465	hypothetical protein IMCC14465_18440
2	52	Galbibacter sp. ck-I2-15	maltose phosphorylase
2	55	Vermamoeba vermiformis	chitinase-like protein cluster A
2	56	Saccharophagus degradans 2-40	hypothetical protein Sde_3003
2	60	uncultured bacterium	cellodextrinase
1	64	Cytophaga hutchinsonii ATCC 33406	b-glucosidase
1	57	Cytophaga hutchinsonii ATCC 33406	glycoside hydrolase family 5
1	53	Cytophaga hutchinsonii ATCC 33406	beta-glycosidase-like protein
1	76	Clostridium novyi NT	4-alpha-glucanotransferase
1	33	Bacillus circulans	unnamed protein product
1	48	Oryza sativa Japonica Group	hypothetical protein OsI_13514
1	34	Clostridium thermocellum	Beta-Glycanase-like
1	49	Lentisphaera araneosa HTCC2155	putative alpha-mannosidase
1	51	Pedobacter sp. BAL39	beta-galactosidase
1	83	Clostridium beijerinckii NCIMB 8052	glycoside hydrolase
1	65	Bacteroides caccae ATCC 43185	hypothetical protein BACCAC_02668
1	68	Roseiflexus castenholzii DSM 13941	4-alpha-glucanotransferase

1	44	Polyporus arcularius	cellobiohydrolaseII
1	51	Nocardiopsis Sp.Strain F96	Endo-Beta-1,3-Glucanase From Alkaliphilic Nocardiopsis Sp.Strain F96
1	61	uncultured bacterium	glycoside hydrolase family 5
1	46	Clostridium acetobutylicum DSM 1731	1,4-beta-N-acetylmuramidase
1	46	Clostridium acetobutylicum DSM 1731	beta-glucosidase
1	65	Herpetosiphon aurantiacus DSM 785	glycoside hydrolase
1	90	Clostridium phytofermentans ISDg	glycoside hydrolase family protein
1	55	Acholeplasma laidlawii PG-8A	glycosyl hydrolase family 3 protein
1	68	Sorangium cellulosum So ce56	xylan 1,4-beta-xylosidase
1	57	Zea mays	beta-glucanase precursor
1	38	Coprococcus eutactus ATCC 27759	hypothetical protein COPEUT_00970
1	44	Eubacterium siraeum DSM 15702	hypothetical protein EUBSIR_00676
1	55	Physcomitrella patens subsp. patens	predicted protein
1	73	Physcomitrella patens subsp. patens	predicted protein
1	59	Exiguobacterium sibiricum 255-15	maltose phosphorylase
1	68	Cellvibrio japonicus Ueda107	endo-1,4-beta glucanase
1	49	Cellvibrio japonicus Ueda107	beta glucanase
1	41	Cellvibrio japonicus Ueda107	beta glucanase
1	65	Phaseolus vulgaris	beta-glucosidase-like protein
1	66	Thermoanaerobacter tengcongensis MB4	trehalose and maltose hydrolase (phosphorylase)
1	40	Geobacillus sp. 70PC53	CelA precursor
1	51	Dictyoglomus turgidum DSM 6724	glycoside hydrolase family protein
1	54	Chloroflexus aggregans DSM 9485	glycoside hydrolase family protein
1	45	Halothermothrix orenii H 168	family 1 glycoside hydrolase6
1	44	bacterium Ellin514	Mannan endo-1,4-beta-mannosidase
1	76	Zea mays	unknown
1	69	Sphingobacterium spiritivorum ATCC 33300	glycoside hydrolase family 2, sugar binding protein

1	44	Mycobacterium kansasii ATCC 12478	Beta-glucosidase
1	53	gamma proteobacterium NOR5-3	glucan 1,4-beta-glucosidase
1	29	Teredinibacter turnerae T7901	glycoside hydrolase family 12 domain-containing protein
1	53	Teredinibacter turnerae T7901	glycoside hydrolase family 5 domain-containing protein
1	66	Dyadobacter fermentans DSM 18053	glycoside hydrolase
1	47	Dyadobacter fermentans DSM 18053	glycoside hydrolase
1	60	Clostridium thermocellum DSM 2360	Beta-glucosidase
1	51	Chitinophaga pinensis DSM 2588	glycoside hydrolase
1	66	Chitinophaga pinensis DSM 2588	glycoside hydrolase
1	55	Brachybacterium faecium DSM 4810	arabinogalactan endo-1,4-beta-galactosidase
1	67	Treponema vincentii ATCC 35580	4-alpha-glucanotransferase
1	63	Dictyoglomus thermophilum	beta-mannanase
1	52	Cronobacter turicensis z3032	6-phospho-beta-glucosidase
1	63	synthetic construct	endoglucanase D variant
1	41	Fibrobacter succinogenes subsp. succinogenes S85	glycoside hydrolase family protein
1	44	Fibrobacter succinogenes subsp. succinogenes S85	mannan endo-1,4-beta-mannosidase
1	50	Fibrobacter succinogenes subsp. succinogenes S85	cellulase
1	66	Fibrobacter succinogenes subsp. succinogenes S85	glycoside hydrolase family protein
1	57	Fibrobacter succinogenes subsp. succinogenes S85	coagulation factor 5/8 type domain-containing protein
1	51	Fibrobacter succinogenes subsp. succinogenes S85	glycoside hydrolase family protein
1	49	uncultured organism	endo-beta-1,3(4)-glucanase
1	68	Streptobacillus moniliformis DSM 12112	4-alpha-glucanotransferase
1	70	Cellulosilyticum ruminicola	galactosidase
1	61	Sulfolobus islandicus L.D.8.5	glycoside hydrolase family protein

1	71	Clostridium hathewayi DSM 13479	4-alpha-glucanotransferase, partial
1	56	Thermoanaerobacter italicus Ab9	glycoside hydrolase family protein
1	64	uncultured organism	putative carbohydrate-active enzyme
1	51	Saccharopolyspora erythraea NRRL 2338	glucan 1,4-beta-glucosidase precursor
1	44	Butyrivibrio fibrisolvens 16/4	Arabinogalactan endo-1,4-beta-galactosidase
1	88	Roseburia intestinalis XB6B4	Alpha-glucosidases, family 31 of glycosyl hydrolases
1	72	Roseburia intestinalis XB6B4	Alpha-glucosidases, family 31 of glycosyl hydrolases
1	64	Ruminococcus sp. SR1/5	4-alpha-glucanotransferase
1	47	Eubacterium siraeum V10Sc8a	Beta-mannanase
1	53	Arabidopsis lyrata subsp. lyrata	glycosyl hydrolase family 3 protein
1	56	Ktedonobacter racemifer DSM 44963	4-alpha-glucanotransferase
1	55	Ktedonobacter racemifer DSM 44963	Alpha-mannosidase
1	67	Bacteroides sp. 3_1_19	periplasmic beta-glucosidase
1	51	Prevotella bryantii B14	putative glycosyl hydrolase
1	59	Amycolatopsis mediterranei S699	beta-glucosidase
1	68	Clostridium ljungdahlii DSM 13528	glycosyl hydrolase
1	59	Enterococcus faecalis str. Symbioflor 1	putative 6-phospho-beta-glucosidase
1	35	Clostridium saccharolyticum WM1	glycoside hydrolase
1	68	Selaginella moellendorffii	hypothetical protein SELMODRAFT_74114
1	81	Clostridium cellulovorans 743B	glycoside hydrolase family 48
1	44	Leadbetterella byssophila DSM 17132	cellulase
1	73	Caldicellulosiruptor owensensis OL	xylan 1,4-beta-xylosidase
1	50	Caldicellulosiruptor owensensis OL	glycoside hydrolase family 5
1	38	Caldicellulosiruptor kronotskyensis 2002	beta-glucuronidase
1	78	Paludibacter propionicigenes WB4	beta-galactosidase
1	96	Paludibacter propionicigenes WB4	glycoside hydrolase
1	83	Paludibacter propionicigenes WB4	glycoside hydrolase

1	45	Bacteroides eggerthii 1_2_48FAA	carbohydrate binding module
1	72	Bacteroides helcogenes P 36-108	N-acylglucosamine 2-epimerase
1	67	Bacteroides helcogenes P 36-108	glycoside hydrolase 3
1	59	Escherichia coli WV_060327	Rhamnogalacturonides degradation protein RhiN
1	68	Treponema phagedenis F0421	putative 4-alpha-glucanotransferase
1	67	Granulicella tundricola MP5ACTX9	alpha-1,2-mannosidase
1	49	uncultured organism	putative carbohydrate-active enzyme
1	55	uncultured organism	putative carbohydrate-active enzyme
1	50	Aureococcus anophagefferens	putative endo-beta-1,3-1,4 glucanase
1	80	Pedobacter saltans DSM 12145	glycoside hydrolase family protein
1	43	Planctomyces brasiliensis DSM 5305	hypothetical protein Plabr_1678
1	51	Koerneria sudhausi	cellulase
1	73	Cellulophaga lytica DSM 7489	cellulase
1	73	Bacteroides salanitronis DSM 18170	alpha-1,2-mannosidase
1	44	Ruminococcus albus 8	beta-mannanase/endoglucanase A domain protein
1	61	Turicibacter sp. HGF1	putative chitinase ChiB1
1	69	Clostridium lentocellum DSM 5427	glycoside hydrolase
1	49	Asticcacaulis biprosthecum C19	periplasmic beta-glucosidase
1	69	Bacteroides clarus YIT 12056	glycosyl hydrolase family 26
1	48	Verrucosispora maris AB-18-032	glycoside hydrolase family protein
1	76	Sphingomonas sp. S17	glycosyl hydrolases 31 family protein
1	81	Treponema brennaborense DSM 12168	4-alpha-glucanotransferase
1	73	Treponema brennaborense DSM 12168	Beta-glucosidase
1	49	Mahella australiensis 50-1 BON	alpha-mannosidase
1	75	Mahella australiensis 50-1 BON	glycoside hydrolase
1	53	Dysgonomonas gadei ATCC BAA-286	hypothetical protein HMPREF9455_00408
1	67	Dysgonomonas gadei ATCC BAA-286	hypothetical protein HMPREF9455_00418

1	94	synthetic construct	putative glycosyl hydrolase
1	74	Methylomonas methanica MC09	Beta-glucosidase
1	63	Treponema azotonutricium ZAS-9	alpha-galactosidase
1	56	Thioalkalimicrobium cyclicum ALM1	4-alpha-glucanotransferase
1	60	Alistipes sp. HGB5	conserved hypothetical protein
1	44	Bacteroides sp. 1_1_30	hypothetical protein HMPREF0127_05086
1	53	Lachnospiraceae bacterium 3_1_57FAA_CT1	hypothetical protein HMPREF0994_03348
1	54	Lachnospiraceae bacterium 1_4_56FAA	hypothetical protein HMPREF0988_02795
1	52	Clostridium acetobutylicum DSM 1731	xylanase
1	74	Spirochaeta caldaria DSM 7334	Licheninase
1	64	Spirochaeta caldaria DSM 7334	cellulase
1	73	Spirochaeta caldaria DSM 7334	glycoside hydrolase family protein
1	51	Termitomyces albuminosus	glycoside hydrolase family 6 protein
1	76	Zobellia galactanivorans	beta-N-acetylglucosaminidase
1	68	Cyclobacterium marinum DSM 745	glycoside hydrolase
1	75	Caldicellulosiruptor lactoaceticus 6A	glycoside hydrolase family protein
1	42	Anaerophaga thermohalophila DSM 12881	alpha-glucuronidase
1	46	Anaerophaga thermohalophila DSM 12881	Mannan endo-1,4-beta-mannosidase
1	35	Verticillium dahliae VdLs.17	exoglucanase-6A
1	50	Lactobacillus ruminis ATCC 27782	maltose phosphorylase
1	70	Thermobacillus composti KWC4	Mannan endo-1,4-beta-mannosidase
1	50	Brachypodium distachyon	PREDICTED: beta-glucosidase 24-like
1	63	Flavobacteriaceae bacterium HQM9	glycoside hydrolase family 16 domain-containing protein
1	66	uncultured bacterium H1_5	GHF3 protein
1	75	Subdoligranulum sp. 4_3_54A2FAA	hypothetical protein HMPREF1032_00735
1	55	Acetivibrio cellulolyticus CD2	glycoside hydrolase

1	51	Acetivibrio cellulolyticus CD2	glycoside hydrolase family protein
1	82	Acetivibrio cellulolyticus CD2	glycoside hydrolase family protein
1	74	Acetivibrio cellulolyticus CD2	glycoside hydrolase family protein
1	76	Anaerophaga sp. HS1	glycoside hydrolase
1	48	Anaerophaga sp. HS1	mannan endo-1,4-beta-mannosidase
1	76	Anaerophaga sp. HS1	family 2 glycoside hydrolase
1	70	Mucilaginibacter paludis DSM 18603	beta-glucosidase
1	43	Mucilaginibacter paludis DSM 18603	Licheninase
1	78	Mucilaginibacter paludis DSM 18603	glycoside hydrolase family protein
1	68	Mucilaginibacter paludis DSM 18603	beta-galactosidase
1	76	Flavobacteriaceae bacterium S85	glycoside hydrolase family protein
1	54	Neocallimastix patriciarum	putative cellulase
1	42	Neocallimastix patriciarum	putative cellulase
1	50	Caldithrix abyssi DSM 13497	glycoside hydrolase family 5
1	51	Clostridium clariflavum DSM 19732	beta-glucosidase-like glycosyl hydrolase
1	36	Paenibacillus cookii	cellulase
1	80	Niabella soli DSM 19437	Xylan 1,4-beta-xylosidase
1	52	Niabella soli DSM 19437	Glycoside hydrolase 97
1	31	Paenibacillus dendritiformis C454	Licheninase
1	61	Paenibacillus dendritiformis C454	Licheninase
1	72	Treponema primitia ZAS-1	xylan 1,4-beta-xylosidase
1	65	Niastella koreensis GR20-10	4-alpha-glucanotransferase
1	67	Paenibacillus sp. Aloe-11	hypothetical protein WG8_2569
1	73	Clostridium sp. BNL1100	beta-xylosidase
1	81	Clostridium sp. BNL1100	glycosyl hydrolase family 11,dockerin-like protein
1	68	Treponema pallidum subsp. pallidum DAL-1	glycosyl hydrolase domain protein
1	41	uncultured bacterium	GH12 cellulase-like protein
1	60	Treponema saccharophilum DSM 2985	4-alpha-glucanotransferase

1	30	Bacteroides sp. D2	hypothetical protein BSGG_4808
1	71	Caldilinea aerophila DSM 14535 = NBRC 104270	putative beta-glucosidase
1	40	Phycisphaera mikurensis NBRC 102666	hypothetical protein PSMK_26070
1	45	Prevotella sp. oral taxon 306 str. F0472	glycosyl hydrolase family 25
1	76	Fibrella aestuarina BUZ 2	glycoside hydrolase family 65 central catalytic
1	80	Imtechella halotolerans K1	beta-N-acetylhexosaminidase
1	61	Coccomyxa subellipsoidea C-169	putative prunasin hydrolase isoform PHA precursor
1	48	Coccomyxa subellipsoidea C-169	Six-hairpin glycosidase
1	31	Ignavibacterium album JCM 16511	Beta-glucanase/beta-glucan synthetase
1	80	Paenibacillus mucilaginosus K02	hypothetical protein B2K_12155
1	58	Paenibacillus mucilaginosus K02	mannan endo-1,4-beta-mannosidase A and B
1	67	Bacillus sp. JS	secreted arabinogalactan oligomer endo-hydrolase
1	53	Streptomyces hygroscopicus subsp. jinggangensis 5008	glucan 1,4-beta-glucosidase
1	75	Actinoplanes sp. SE50/110	arabinofuranosidase
1	55	Actinoplanes sp. SE50/110	beta-glucosidase
1	62	Haemophilus haemolyticus HK386	4-alpha-glucanotransferase
1	38	Nannochloropsis gaditana CCMP526	glycoside hydrolase family 8
1	73	Solitalea canadensis DSM 3403	beta-glucosidase-like glycosyl hydrolase
1	73	Solitalea canadensis DSM 3403	beta-glucosidase-like glycosyl hydrolase
1	56	Emticicia oligotrophica DSM 17448	glycoside hydrolase family 65 central catalytic
1	52	Cellvibrio sp. BR	endo-beta-1,3(4)-glucanase
1	36	Rhodanobacter fulvus Jip2	glycosyl hydrolase family 32 protein
1	35	Streptococcus suis ST1	glycoside hydrolase family protein
1	47	Opitutaceae bacterium TAV1	putative glycoside hydrolase
1	80	Paenibacillus peoriae KCTC 3763	beta-glucosidase/6-phospho-beta- glucosidase/beta-galactosidase
1	40	Thermoanaerobacterium saccharolyticum JW/SL-YS485	hypothetical protein Tsac_2300

1	69	Saccharophagus sp. Myt-1	cellulase
1	42	Coniophora puteana RWD-64-598 SS2	glycoside hydrolase family 25 protein
1	50	Bacteroides cellulosilyticus CL02T12C19	hypothetical protein HMPREF1062_03427
1	71	Bacteroides cellulosilyticus CL02T12C19	hypothetical protein HMPREF1072_00265
1	41	Fibrisoma limi BUZ 3	chitinase
1	35	Fomitiporia mediterranea MF3/22	cellulase CEL6B
1	70	Bacteroides salyersiae CL02T12C01	hypothetical protein HMPREF1071_00445
1	61	Bacteroides salyersiae CL02T12C01	hypothetical protein HMPREF1071_03465
1	60	Capnocytophaga sp. oral taxon 335 str. F0486	glycosyl hydrolase, family 57
1	44	Leptosphaeria maculans JN3	hypothetical protein LEMA_P018790.1
1	76	Melioribacter roseus P3M	beta-glucanase precursor
1	79	Melioribacter roseus P3M	glycoside hydrolase family 43
1	62	Melioribacter roseus P3M	glucosyl hydrolase family protein
1	74	Chryseobacterium sp. CF314	alpha-glucuronidase
1	60	Flavobacterium sp. CF136	arabinogalactan endo-1,4-beta-galactosidase
1	71	Flavobacterium sp. CF136	beta-xylosidase
1	61	Streptococcus ratti FA-1 = DSM 20564	putative endoglucanase precursor
1	51	Aquimarina agarilytica ZC1	glycoside hydrolase
1	51	Aquimarina agarilytica ZC1	cellulase
1	41	Alistipes sp. JC136	maltose phosphorylase
1	65	Barnesiella intestinihominis YIT 11860	hypothetical protein HMPREF9448_02298
1	58	uncultured bacterium	hypothetical protein ACD_76C00045G0001
1	53	uncultured bacterium	hypothetical protein ACD_38C00018G0010
1	69	uncultured bacterium	hypothetical protein ACD_34C00276G0003
1	75	uncultured bacterium	hypothetical protein ACD_20C00317G0008

1	84	Bacillus sp. HYC-10	alpha-galactosidase
1	51	Piromyces rhizinflatus	exocellobiohydrolase precursor
1	60	uncultured bacterium	endo-1,4-beta-xylanase precursor
1	56	uncultured bacterium	putative cellulase
1	53	Piromyces rhizinflatus	exocellobiohydrolase precursor Cbh6
1	50	Fusarium sp. IFO 7772	hypothetical protein similar to beta-D-galactosidase
1	39	Bacillus cereus E33L	chitosanase
1	61	Aspergillus nidulans FGSC A4	hypothetical protein AN7413.2
1	50	Hahella chejuensis KCTC 2396	cellulase
1	56	Saccharophagus degradans 2-40	regulatory protein, LacI
1	56	Saccharophagus degradans 2-40	cellulase
1	52	Deinococcus geothermalis DSM 11300	glycoside hydrolase family protein

5.15. Discussion

This chapter describes the analysis of an environmental metatranscriptome, in the form of a dataset produced from Illumina paired-end sequencing. The data were thoroughly screened for low-quality reads and rRNA sequences all of which were removed before phylogenetic and functional analysis of the data. The non rRNA putative protein-coding sequences were analysed to gain an understanding of the functions being carried out by the community, and the identity of the active participants. In addition to an overall phylogenetic and functional classification using the MEGAN metagenome analyser software and the MG-RAST webserver the data was specifically screened for reads encoding glycoside hydrolase enzymes. Sensitive searching against the Pfam database identified a total of 503 probable glycoside hydrolase sequences.

5.14.1 Technical constraints and Data processing challenges

The choice of sequencing platform made here was an operational consideration. Pyrosequencing, with its poor performance when dealing with sequences containing long homopolymer repeats had been used in the first instance as described in chapter 4 and the long polyA tail regions present in the sample had indeed resulted in truncated and deprecated read data. Although the inclusion of polyA tails in the final cDNA library can be minimised with modifications of the sample, preparation protocol sections consisting of multiple polyA sequences will likely remain on the ends of some sequences and might, therefore, result in a proportion of low quality reads. At initial conception of the project, Illumina sequences was delivering much shorter reads than can currently be achieved with the technology but after an initial failure to obtain good quality data with 454 pyrosequencing improvements in Illumina read length rendered it a more attractive option.

With this in mind, it was decided that Illumina sequencing on the Miseq platform offered a safer option since there was no risk of errors arising from homopolymer repeats at using this technology. The shorter read lengths produced by this technology were offset by more a consistent read length, avoiding excessively short sequences, the ability to create longer reads by pairing of overlapping mate pairs on the one hand and the fact that the Miseq would produce a much greater read depth. Although there were solutions available to the issues encountered when performing pyrosequencing on the type of sample being investigated here, Illumina sequencing still remained an attractive option. This was

due in part to issues being reported with the local 454 machine which raised concerns that the quality of data that would be obtained might be lower anyway. Additionally, it was possible that after polyA removal a proportion of the sequences would still contain relatively long polyA sequences. This is because the length of the tail and the proportion of the polyA regions which would be removed would be to an extent determined randomly by the location at which the primer used at the beginning of the MessageAmp protocol bound to the sequence. There would therefore be an expectation that on a pyrosequencing run a proportion of reads would fail due to homopolymer-associated errors and this was coupled with concerns about the overall performance of the machine itself. Sequencing of cDNA libraries produced from MessageAmp amplified RNA with enzymatically removed polyA tails has been achieved successfully with pyrosequencing but has been found to consistently yield fewer reads than sequencing of genomic DNA libraries from the same samples and it has been suggested that Adenosine-rich sequences remaining after enzymatic digestion may be responsible for this (Shi *et al*. 2011). It was therefore decided ultimately that the Miseq represented a much safer choice for producing good quality, reliable data.

The choice of sequencing the community metatranscriptome on the Illumina platform was constrained by circumstance and resulted in a successful sequencing run producing a large amount of high quality data. This is one of only a very few metatranscriptomic sequencing efforts carried out using Illumina sequencing as most groups have, thus far, preferred to opt for the longer read length of 454. Ultimately, comparison of the data output from the Miseq sequencing run, and the average read-length achieved after merging of overlapping paired-end reads demonstrated that the Illumina platform was a highly effective option. The output was far larger than datasets produced by 454 sequencing, which is a benefit in that a greater amount of data is generated but also a drawback in that the sheer volume of the Miseq data became itself a challenge and a barrier to straightforward analysis. Most tools for the analysis of metagenomic or metatranscriptomic data produced by high throughput sequencing were developed when 454 datasets were beginning to emerge as a standard feature of molecular microbial ecology. Datasets containing millions of Illumina reads are not well catered for and the amount of sequence data requires a great deal of computational power to process. Even with access to powerful computational resources traditional analysis methods, such as blast, are not well suited to this type of data and processing times become protracted.

Quality control was performed with relatively high stringency. Both quality screening and rRNA screening settings were at conservative levels. The nature of

the analysis that was to be performed on the data dictated that it was important to ensure that only sequences of high quality, free from errors, were passed to later stages of the analysis pipeline in order to avoid misannotations. Quality control using the Prinseq program revealed that the initial concern about the presence of a large number of sequences with long polyA tails in the dataset may have been exaggerated as only a small number of sequences (less than 2%) had polyA regions at the 5' or 3' end and many of these were in any case less than ten nucleotides in length. What quality control did reveal however was a huge number of artificial duplicate reads, which comprised approximately 70% of the total. Artificial duplicates are a feature of any metagenomic study (Gomez-Alvarez *et al*. 2009) but the proportion here seems to be very high. There were also a large number of rRNA sequences in the data; rRNA comprised approximately 71% of the sequences after initial quality filtering. These two features may be related, as naturally highly abundant rRNA sequences might have become artificially increased during processing of the sample. This could conceivably have taken place during library preparation before sequencing or during the MessageAmp amplification steps, or as a result of multiple steps exerting a synergistic effect. Without extensive experimentation it will not be possible to answer this question fully but given the various enzyme-dependant steps involved in the production of a dscDNA library from an environmental RNA sample there is plenty of scope for sequences that were already present in high numbers to have been preferentially amplified at multiple points.

Although an rRNA removal step had been carried out, a great deal of rRNA remained in the sample that was submitted for sequencing and was removed from the dataset before functional analysis. Removal of rRNA was carried out before amplification with MessageAmp in an attempt to limit excessive amplification of abundant rRNA in the sample but despite this a majority fraction of the dataset obtained consisted of rRNA reads. It is well known that rRNA is the overwhelmingly abundant type of RNA molecule in cells and metatranscriptomic studies are always faced with rRNA removal as a major technical challenge. Attempts to remove rRNA before metatranscriptomic sequencing tend to result in 30-80% of the dataset remaining as ribosomal in origin despite the best efforts of researchers and the use of multiple removal strategies (Stewart *et al*., 2010, Shi *et al*., 2011) although one study did report an mRNA enrichment of 99%, an exceptional occurrence which has not been repeated since (Gilbert *et al*., 2008). A metatranscriptomic sequencing project where the rRNA content comprises 71% of the total reads (after removal or poor quality sequences and artificial duplicates) is not an unexpected result and after

rRNA removal nearly one million sequences representing putative protein-coding mRNA reads remained for analysis. Indeed, the goal of rRNA removal is not a complete avoidance of ribosomal sequences in the final dataset, as this is unrealistic, but instead to maximise the number of cDNA reads originating from mRNA.

The MG-RAST pipeline can predict both functional protein and rRNA features, and some of the annotations produced after processing the metatranscriptome did call into question how complete the rRNA removal step using Ribopicker had been. The identification of 21, 325 rRNA features by MG-RAST suggests that screening and removal of rRNA sequences had not been entirely efficient. Given the large proportion of rRNA sequences present in the initial dataset it is likely that some would make it into the functional screening, especially since shorter Illumina reads covering hypervariable sections of the 16S or 18S rRNA gene of an organism poorly or not at all represented in the databases might slip through a screening process which relies on homology and sequence identity percentages. This is unfortunate but unavoidable and when working with datasets of millions of sequences there is always likely to be some level of misannotation. The Blastx search of the putative mRNA reads did return a large number of no hit results, and searching against a non-redundant protein database with an rRNA query sequence should ideally result in a no hit report for some of the sequences It is likely that the presence of a few rRNA reads in the functional analysis has had a minimal impact on the results, and the conclusions drawn.

Assembly of the data was not deemed feasible after attempts with the velvet extensions metavelvet, and oases failed to produce satisfactory results. This is not surprising as the data was extremely uneven and would present a significant challenge for an assembler. Given the unevenness of the blast results, assembly would realistically be limited to a few contigs put together from the high coverage zones and would not be particularly useful even if achievable. Merging paired-end reads was instead performed to generate longer reads and facilitate gene calling.

5.14.2 Functional Analysis: strengths and limitations.

A metatranscriptome is in many ways highly subjective and the genes picked up through metatranscriptomic sequencing will depend on the conditions of the environment and the effect of those conditions on the resident microbes. A metatranscriptome is therefore never able to reveal the full metabolic potential of an environment, but does provide a specific snapshot of the gene expression of a community at a particular time point in response to specific conditions. Here, this specificity was harnessed to attempt to identify cellulase genes, expressed by

members of a microbial community resident in an anaerobic freshwater lake sediment where a crystalline cellulose substance had been introduced. The hypothesis was that within the biofilm formed on the crystalline cellulosic material, specialist cellulose degrading organisms would be present and in elevated numbers as the transcription levels for these genes would be increased as the bacteria expressed the proteins needed for utilisation of the cellulose as a carbon source.

Phylogenetic breakdowns of metatranscriptomic reads reflect the transcriptional activity of individual organisms rather than absolute species presence, but is a good indication of which organisms are the active members of the community in the anoxic sediment. The phylum-level distribution of reads is an unsurprising pattern. Proteobacteria, Bacteroidetes, Firmicutes and Actinobacteria are diverse assemblages and their representatives are found in high numbers in most environments including soils (Urich *et al*. 2008, Tveit *et al*. 2013), the rumen (Brulc *et al*. 2009) and marine sediments (Wang *et al*. 2012).

Given how the samples were processed and prepared for sequencing, there is scope for bias to have been introduced due to preferential amplification during one of the enzyme-dependant steps. However, since the data were produced from a single library preparation and sequencing run any bias should be broadly similar across all the groups and there is no reason why one organism should have been disproportionately affected by the treatments.

There were differences in the phylogenetic profiles of the community as determined by MEGAN and MG-RAST analysis of the data but the overall pattern was similar. Some Phyla were highly abundant. Some (e.g. the Verrucomicrobia) were found represent at 1% of the total reads by both analyses and presumably contribute some level of activity to the microbial community. Some (e.g. the Fibrobacteres which were 0.12% of the total by MEGAN classification and 1.59% by MG-RAST) had different abundances determined by the two analyses, which suggests that they still might have a contribution to overall community activity. Several phyla were determined to constitute less than 1% of the total reads and these groups possibly have very little functional activity in the anoxic sediment. Cells from the water column could end up trapped in the sediment on occasion and cells from other environments could sometimes be washed in the lake and persist in the sediment for some time, but not as a major part of the microbial ecosystem. The presence of some phyla is therefore likely to be incidental. Some organisms may have become introduced to the sample during the harvesting step as the cellulose baits were pulled up through the water column. This would explain the apparent detection of low numbers of reads corresponding to Cyanobacteria from the

sediment as these organisms are normally restricted to the water column. Alternatively, the reads classified as Cyanobacteria could be from related organisms with conserved proteins.

The metatranscriptome analysed here was associated with a cellulose bait in order to enrich for organisms able to colonise and utilise the cellulose as a carbon source. However, these cellulose baits were placed directly into a lake sediment with an established population of microbes. There will be a great deal of biology taking place in a lake sediment and this sequencing project will have detected a great deal of general metabolic activity from non-cellulose degrading organisms. Low number of, e.g., *Fibrobacter* sequences in the overall data means very little in terms of the particular activity of members of this phylum against cellulose within this biofilm community or ecosystem. If reads associated with *Fibrobacter* or close relatives of this organism are indicative of its presence and active gene expression, then it may be localised specifically to the cotton bait in large numbers. Reads of other organisms could represent species which are present in sediment material covering the cotton baits and which have no specific association with the cellulose baits at all.

MG-RAST mapped 193 sequences to the *Fibrobacter succinogenes* S85 genome and these reads were mapped to general housekeeping genes such as transcription and translation factors in addition to gene from GH families, which may indicate active expression of genes involved with polysaccharide breakdown. In any case, it would appear that relatives of this lineage, once described as a gut-specific symbiotic degrader of cellulosic biomass in ruminants, exist in other environments and are metabolically active. In addition to their detection in lake sediment here they are apparently also involved in cellulose breakdown in landfill sites (McDonald *et al*., 2012) and as all cultured members thus far have proven to be cellulose degraders (Ransom-Jones *et al*., 2012) it is almost certainly the case that the *Fibrobacter* or *Fibrobacter* relatives detected in the metatranscriptome are involved in this process.

One functional feature that stood out from the SEED subsystem classification by MG-RAST was a large number of phage-related sequences. MG-RAST mapped a large number of reads to a Enterobacteria Phage genome and the discovery of a phage as contributing to the ecology of microbes in a lake sediment is somewhat consistent with studies of other communities. Other studies have discovered that phage predation is an important factor in the ecology of aquatic populations of bacteria (Ghai *et al*. 2010, Lauro *et al*. 2010). The mapping of a large number of genes to a particular phage genome might indicate the presence of not Phage PhiX174 in particular but a close relative or multiple relatives, with conserved genes. The Enterobacteriaceae constituted 2% of the total bacterial population and given

the high level of phage activity detected it is possible that the host range of the phage detected in the metatranscriptome is in fact more widespread thoughout the Gammaproteobacteria.

Functional and phylogenetic analysis of the metatranscriptome revealed a diverse and active microbial community. Going beyond assignment of reads to broad groups, blast searching revealed that the strongest matches of a large number of the non-rRNA reads in the dataset were in fact to hypothetical and predicted proteins whose function was unknown and in some cases to uncultured, uncharacterised organisms. This reveals a limitation imposed by the data available in the database. Defined functions require experiment evidence or, failing that, careful, precise and highly accurate *in silico* prediction on the level of individual sequences and has been done for only a few organisms.

Use of other searching methods, such as searching against HMM models of Pfam families which are curated and their function well known, provide more useful strategies than using blast for this type of analysis. Searching putative protein sequences against HMM models is slower than blast, and can only be done with subsets of large datasets. This approach is also more sensitive and hits to an HMM model provide a much more definite idea of the identity and function of the query sequence than a blast hit to a “hypothetical protein”.

Using a stringent, sensitive search against the Pfam databases of HMM models identified 503 candidate glycoside hydrolase genes from the metatranscriptome. Some of these sequences had no known match in the NCBI NR database, and none of the sequences had a match with an identity above 85%. It is possible that relaxing of the search criteria may have identified hits for some of the unmatched sequences but at the cost of obtaining matches with a greater likelihood of being co-incidental. It seems likely that the glycoside hydrolase sequences detected here may contain examples of novelty, if not in terms of catalytic activity and new categories of glycoside hydrolases then perhaps in terms of enzymes with distinct sequences and structures which confer higher activity at lower temperatures, pH tolerance and other traits potentially useful in an industrial setting. Other studies utilising cloning of metagenomic DNA and functional screening have typically identified glycoside hydrolase sequences successfully, although the hit rate from such studies has always been low as tens of thousands of clones generally yield between 10 and 30 positives even when targeting environments where such enzymes are prevalent such as the rumen (Ferrer *et al*., 2012) or enriching for cellulose degrading organisms by supplementing natural soil with avicel (Takasaki *et al*., 2013). The only benefit of a clone-and-screen approach to searching for

glycoside hydrolase enzymes in a metagenome is that when a positive clone is identified the protein can be produced in high quantities and further characterised biochemically, although this benefit is a desirable one. High throughput sequencing will always yield far greater numbers of glycoside hydrolase sequences, often in the hundreds when the sampling source is a bioreactor harbouring a cellulose degrading community (Xia *et al.,* 2013) or a bovine rumen (Hess *et al*., 2011).

Chapter 6: Analysis of a lake sediment metagenome and comparison with the metatranscriptome.

6.1 Background

The metagenome was sequenced as an accompaniment to the metatranscriptome dataset. While the metatranscriptomic dataset (chapter 5) provides information on genes expressed in the colonised cellulose biofilm, the metagenome would provide a much more general overview of the gene content of the community. Genes encoded within the metagenome of the community with polysaccharide degrading function will be sequenced and this is the usual approach to gene disovery in this field (Prakash & Taylor, 2012) but actively transcribed genes detected via the metatranscriptomic approach can be taken to have a much more assured role in the metabolic activity of the community. The metagenomic dataset can however inform analysis of the metatranscriptome and assist in its interpretation Using a comparative analysis of both datasets, specific genes or more general functions represented by KOs, SEED subsystems or COG categories that appear in both datasets might be found in a much higher proportion in the metatranscriptome which would be a strong hint that it represented the metabolic activity of the community.

The samples from which nucleic acid material was extracted to generate material for sequencing of the metatranscriptome and metagenome were not obtained concurrently. This is something of a limitation as the datasets are therefore derived from different populations and makes for an imperfect comparison. In both cases, the samples of nucleic acids were pooled from multiple extractions from colonised baits which were distributed randomly throughout the lake in any case. This chapter therefore presents an overall assessment of the microbiota found in the sediment of Esthwaite Water rather than a specific analysis of a single location.

The difference is the sizes of the datasets also precluded a comparison based on absolute numbers. Abundances of specific groups and functional categories are therefore presented in percentage terms.

6.1.1 Generation of a DNA sampled for sequencing

Total community DNA was extracted from colonised cotton baits that had been deposited into the settlement of Esthwaite Water and stationed there for 10 weeks (July - October 2012) using the method of Griffiths *et al*. (2000). This DNA was cleaned of chemical contamination and RNA as described in chapter 2. The

purified DNA sample was submitted to the Centre for Genomic Research for sequencing on the Illumina Miseq platform, using 250x250 paired-end sequencing.

6.1.1 Data analysis considerations

This metagenomic dataset is larger than the metatranscriptome in terms of both numbers of sequences and sequence length in comparison to the data in chapter 5 and this fact has certain implications for data analysis. The metatranscriptome raw reads occupied 5.04GB, and after quality filtering a large number of the reads, the final file used for functional analysis contained just under one million sequences. The metagenome consisted initially of 11.81 GB of sequence data and the majority of these reads passed QC analysis as the proportion of artificial duplicates was much lower. The sheer size of the metagenome dataset and the number of reads it contained was, while a valuable source of information, challenging to analyse.

Previous approaches using blast searching incurred long processing times and were not practical options for dealing with a dataset of this size. For functional analysis, MG-RAST was selected as being the most suitable method of processing millions of metagenome reads and delivering a comparison with the metatranscriptomic dataset. The fact that the metatranscriptome had already been uploaded and annotated by MG-RAST was an added convenience factor.

6.2 Sequencing output and analysis with Prinseq

Sequencing output was subjected to an initial quality control by the CGR and consisted of three files in fastq format, representing forward and reverse reads and a file of singletons where one of a read pair had failed QC. There were ca. 11.5 million paired-end sequences and ca. 73, 000 singletons.

Before undertaking and processing of the data, the Prinseq program was used to assess the quality of the sequences in the data files. Table 6.1 provides a summary of basic descriptive statistics for the paired-end and singleton files. Unlike the metatranscriptome (where the raw data consisted of reads of invariably 150bp) the initial QC step performed on the metagenome had resulted in some reads being trimmed. Average reads lengths of the metagenomic sequences post initial QC are listed in table 6.1.

Table 6.1 Descriptive statistics for the metagenome dataset determined by analysis with Prinseq

Parameter	File 1 (Forward reads)	File 2 (Reverse reads)	Singletons

Number of Sequences	11,518,566	11,518,566	73,643
Average Length (bp)	227.47 ± 34.58	209.05 ± 44.34	205.40 ± 60.33
Mean GC content	57.99 ± 10.57 %	58.01 ± 10.67 %	59.48 ± 9.98 %
Ambiguity (Bases read as N)	0.13 %	0.51%	4.35 %
PolyA tails present	0.10 %	0.10 %	1.16%
Artificial duplicates	0.39 %	1.15 %	0.79 %

Only a very small percentage of reads contained ambiguity or a polyA tail; the majority of the ambiguities had only a single N read and the length of the polyA tails for most sequences was < 10 bp which probably represents genuine coding sequences, although there were a very small number of reads with large A repeats (the longest was reported at 242 bp). There was also a drop in base quality in the reverse sequence file where the last 10 bases of many reads had a quality score below 20. The proportion of artificial duplicates was much lower than had been previously found for the metatranscriptome where duplicates reads had accounted for approximately 70% of the total output (section 5.4).

Generally the dataset was of very high quality and it was decided to perform a merging step to combine overlapping paired end reads into single contiguous sequences before any further processing of the data.

6.2 Paired-end read assembly with Pandaseq

The Pandaseq Paired-end assembler (Masella *et al*., 2012) was used to perform merging of overlapping read pairs in a similar process to the one carried out using the FLASH utility for the metatranscriptome. Pandaseq is a computationally faster and more sophisticated variant on the functionality of the FLASH program, better suited to handling large datasets.

The full ca. 11.5 million reads were processed with Pandaseq and 10.8 million reads were successfully merged to form longer sequences. This represents a pairing rate of 93.9%. The paired reads were generated from Pandaseq in fastq format with quality information retained and then subjected to a second QC filtering process with Prinseq.

6.3 Quality Control

Although the CGR performed an initial quality screen of the metagenomic data, the reads assembled with Pandaseq were subject to a thorough QC analysis

using Prinseq as before. Given the characteristics of the data as summarised in Table 6.1, the following filtering parameters were implemented:

- Removal of artificial duplicate sequences
- Removal of sequences with more than 1 ambiguous base (N read)
- Removal of sequences of low complexity, defined as an entropy score below 60
- polyA regions of > 5 nt at the 5' or 3' end of a read were trimmed
- Bases with quality scores < 20 at the 5' or 3 ' end of a read ere trimmed
- Sequences shorter than 100 bp were removed

This is effectively identical to the processing of the metatranscriptome dataset except there was no removal of sequences with an average quality score below 20, as sequences with a low average quality score had been removed in the initial QC step performed by the CGR. Sequences that had been trimmed to < 100 bp by quality processing were also removed. These filtering parameters were applied to the data and the filtered output files were used in the downstream analysis.

6.5 Comparative analysis of the metagenome and metatranscriptome of colonised cotton from the sediment of Esthwaite Water using MG-RAST

Data from the MG-RAST analysis of the metatranscriptome and metagenome datasets was retrieved from the website and used to generate comparative plots. The phylogenetic and functional similarities and differences between the two datasets were explored. In all cases, numbers of reads of specific phylogenetic groups or functional categories were expressed at a percentage of the total reads of the corresponding higher level classification i.e the abundance of a particular bacterial phylum was expressed as a percentage of the total number of bacteria.

6.5.1 A Phylogenetic comparison between the datasets.

To examine differences in representation, the two datasets were compared at various taxanomic levels. At the domain level, presented in Fig 1, the dominance of bacteria in the lake sediment was even greater in the metagenome than the metatranscriptome. The bacteria were the overwhelmingly abundant domain, comprising 97.1% of all reads assigned at domain level. The Eukaryotes and viruses were clearly the second and third most abundant domains in the metatranscriptome but only represent very small fractions of the metagenomic reads. The Archaea are actually the second most abundant group in the metagenome, constituting a slightly higher fraction of the reads than in the metatranscriptome but ultimately all domains

are hugely outnumber by the Bacteria which represents most of the metagenomic diversity.

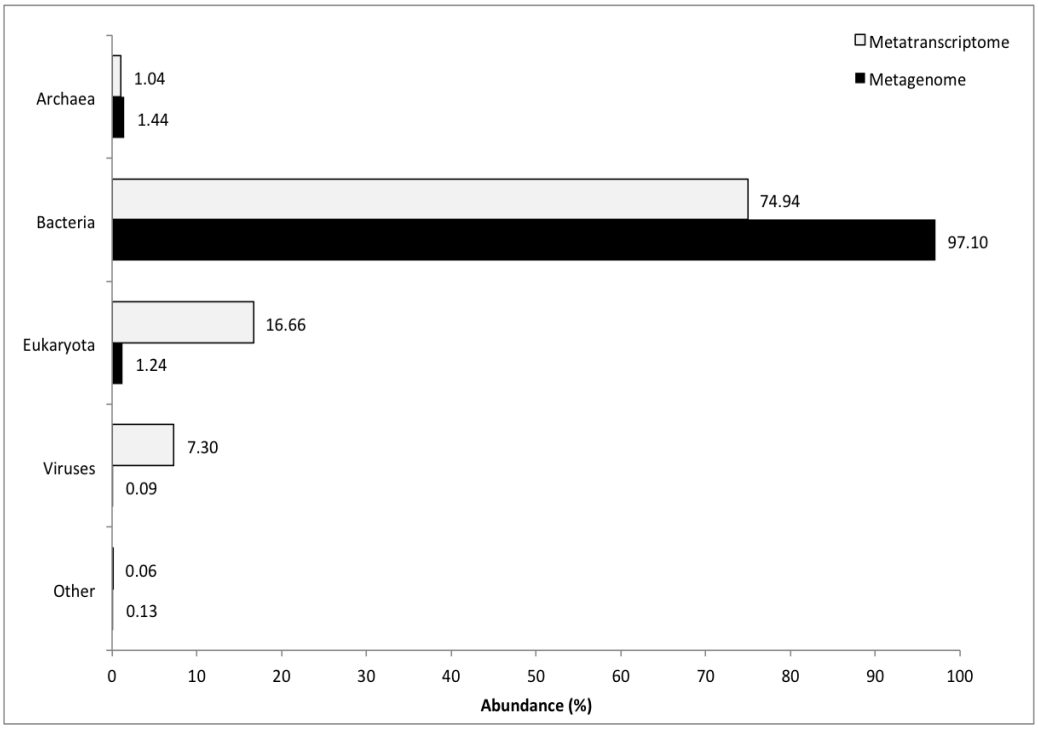


Fig 6.1 Domain-level abundance of reads in the metatranscriptome and metagenome datasets.

6.5.1.1 Bacterial Phylum Diversity

At the bacterial phylum level, the three most abundant phyla are the same, but their abundance with respect to one another differed (Fig. 6.2). The Proteobacteria comprised over 50% of the metagenome bacterial sequences but only 22% of the metatransciptome reads. The Bacteroidetes and Firmicutes were much more relatively abundant in the metatranscriptome and the reduction in relative percentage was particularly acute for the Bacteroidetes. Members of these three phyla are nevertheless predominant in both datasets. The Actinobacteria had the fourth largest proportion of reads assigned by metatranscriptomic analysis, but in the metagenome there are slightly higher numbers of Verrucomicrobia. For these two phyla it appears by metagenomic analysis that they may be present in roughly similar numbers, but the Actinobacteria appear to be the more active.

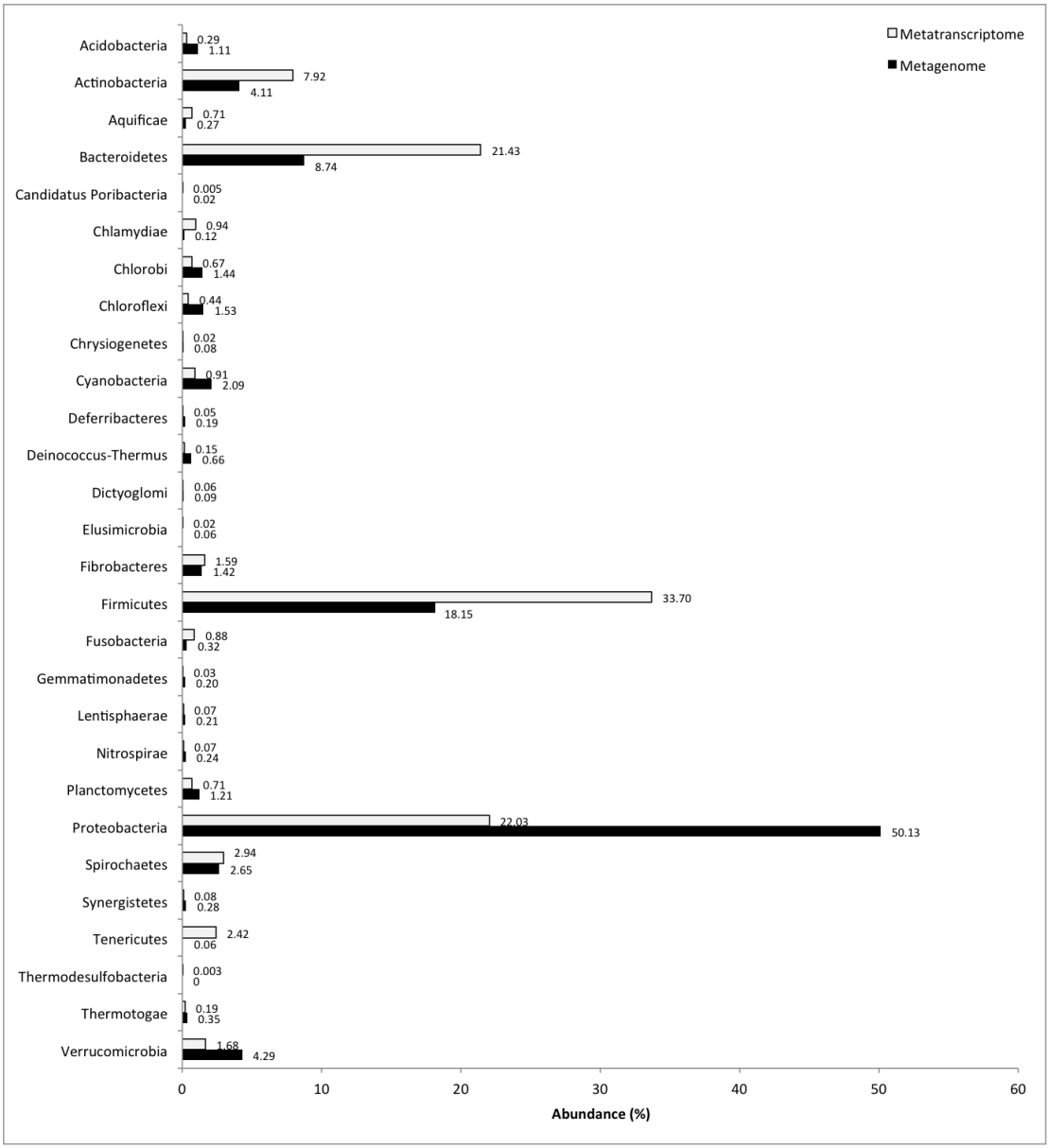


Figure 6.2 Comparison of phylum-level assignment of metatranscriptomic and metagenomic bacterial reads as determined by the MG-RAST pipeline.

6.5.1.2 Eukaryotic Phylum Diversity

Figure 6.3 presents the phylum-level distribution of the eukaryotic reads for all phyla that constituted more than 1% of the total Eukaryotic reads in either the metatranscriptome or metagenome dataset. Many of the reads detected were probably incidental to the cellulose baits and any cellulose degradation that was occurring. The Chordata represented approximately 14% of the Eukaryotic reads in the metagenome and only 1.5% in the case of the metatranscriptome. These reads are probably derived from fish and amphibian faeces and shed cells present in the sediment which would be normally present in such an environment. Additionally, some sequences of human origin might represent contamination. The Streptophyta phylum represents land plants and some algae and their presence in a lake is also understandable as is the presence of Chlorophyta (green algae). The phylum Platyhelminithes (flatworms) was especially well represented in the metatranscriptome but not the metagenome, perhaps reflection a local abundance of such organisms at the time of sampling for the metatranscriptome but not for the metagenome. All members of the Cnidaria are acquatic organisms and Arthropoda, Apicomplexa (protists), Bacillariophyta (diatoms) all have members likely to be found in a lake environment. Ascomycota and Basidiomycota are fungal phyla, and found ubuiquitously in soil and aquatic environments (Orsi *et al*., 2013; Zumsteg *et al*., 2012). Members of the Ascomycota in particular do exhibit polysaccharide degrading activities (*T. reesei* is a member of this phylum) but these were not well represented in the metatranscriptome and are most likely not contributing to cellulose degradation to any great extent.

Not listed in figure 6.3 are the Neocallimastigaceae which were 0.06% of all Eukaryotic reads in the metatranscriptome and 0.3% in the metagenome. Their negligible presence is noteworthy as, like the fibrobacters, they are highly active cellulose degraders in the ruminant gut which might have environmental cousins (Van Dyke & McCarthy, 2002). It would appear that at least in this environment there are very few representative of the phylum present.

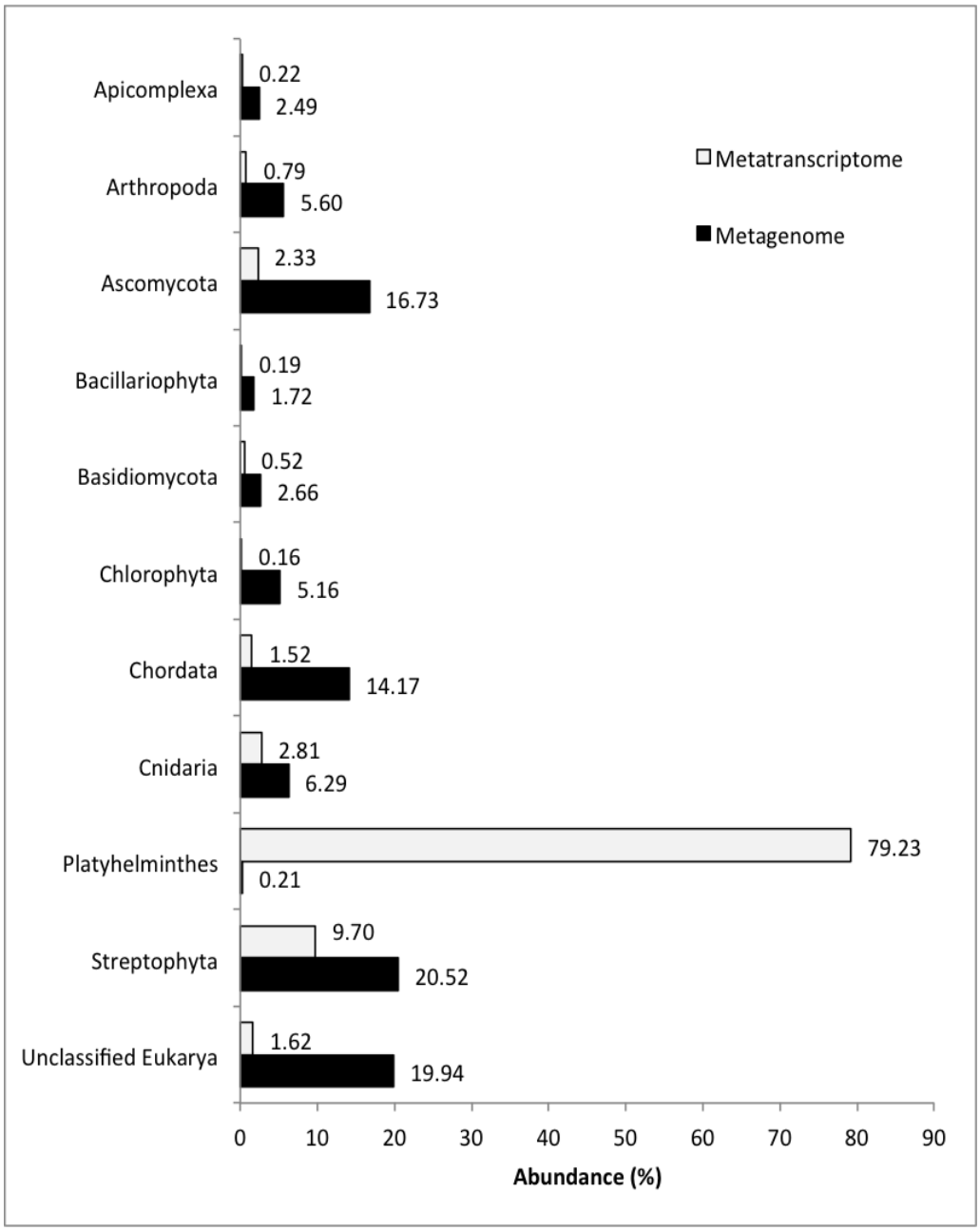


Figure 6.3 Eukaryotic Phylum-level read assignments for the metatranscriptome and metagenome for phyla contributing more 1% of total Eukaryotic reads in at least one of the datasets.

6.5.2 Phylogenetic analysis at genus level.

At the genus level, MG-RAST assigned reads to 1795 genera for the metatranscriptome and 2469 in the case of the metagenome. An exhaustive phylogenetic analysis below phylum level can become difficult and complicated, but examining some specific lineages associated with cellulose degrading capacity does provide some potentially interesting biological insights.

Table 6.2 lists cellulolytic species, found in both the metatranscriptome and metagenome datasets and contrasts their prevalence in both. It appears that while some species (*Cytophaga hutchinsonii* and *Fibrobacter spp*.) are more prevalent in the metatranscriptome, suggesting they are highly active members of the cellulose bait colonising community, some species (*Ruminococcus albus* and *Clostridium cellulolyticum*) actually appear to be relatively more numerous in the metagenome. This observation is especially interesting as *C. hutchinsonii* and *Fibrobacter spp*. have been hypothesised to attack cellulose using an as yet uncharacterised mechanism. *R. albus* and *C. Cellulolyticum* on the other hand utilise cellulosomes to break down cellulose, a mechanism which has been comparatively better studied. It is possible that organisms colonising and actively degrading the cotton baits are expressing previously unknown catalytic mechanisms to do so.

Table 6.2 Porportion of reads to which known cellulolytic species were assigned in the two datasets.

Species Name	**Metatranscriptome**		**Metagenome**	
	Proportion of total phylum reads (%)	Proportion of total bacteria reads (%)	Proportion of total phylum reads (%)	Proportion of total bacteria reads (%)
Cytophaga hutchinsonii	2	0.6	0.3	0.3
Ruminococcus albus	0.2	0.07	2	0.4
Clostridium cellulolyticum	0.8	0.2	3	0.6
*Fibrobacter spp**	100	1	100	1

*Reads corresponding to the Fibrobacteres phylum at species level, representing either *F. succinogenes* or *F. intestinalis*, were grouped together for this comparison.

6.5.3 A functional comparison

Functional comparisons were made between the two datasets based on categorisations of reads to SEED subsystems, KOs and COG categories. This functional categorisation was intended to pinpoint specific metabolic activities that might be better represented by the metatranscriptome, indicating that they are expressed by the biofilm community.

A comparison of the representation of SEED subsystems is presented in Fig 6.3. For most subsystems the level of assignment of reads was broadly similar in both datasets but there is a clear discrepancy in the case of the Phages, Prophages, Transposable elements, Plasmids Subsystem, where a far greater proportion of all SEED categorised reads were assigned for the metatranscriptome. Although some phage have RNA genomes and would not be detected via isolation of DNA, but a large number of reads from the metatranscriptome were found to originate from gene expression of Phage PhiX174 which has an ssDNA genome and should be detected through DNA extraction. The difference seen in this case might be due to temporal variation in phage activity between the two sampling points.

Comparisons between COG categories and KEGG Orthologies are presented in Figs 6.4 and 6.5 and again there are broad similarities between the two datasets and only small differences in the numbers of reads assigned to most categories. Two COG categories (*Translation, ribosomal structure & biogenesis*, and *Energy production and conversion*) appear to be better represented by the metatranscriptome, which suggests these pathways are in heavy use by the microbial community.

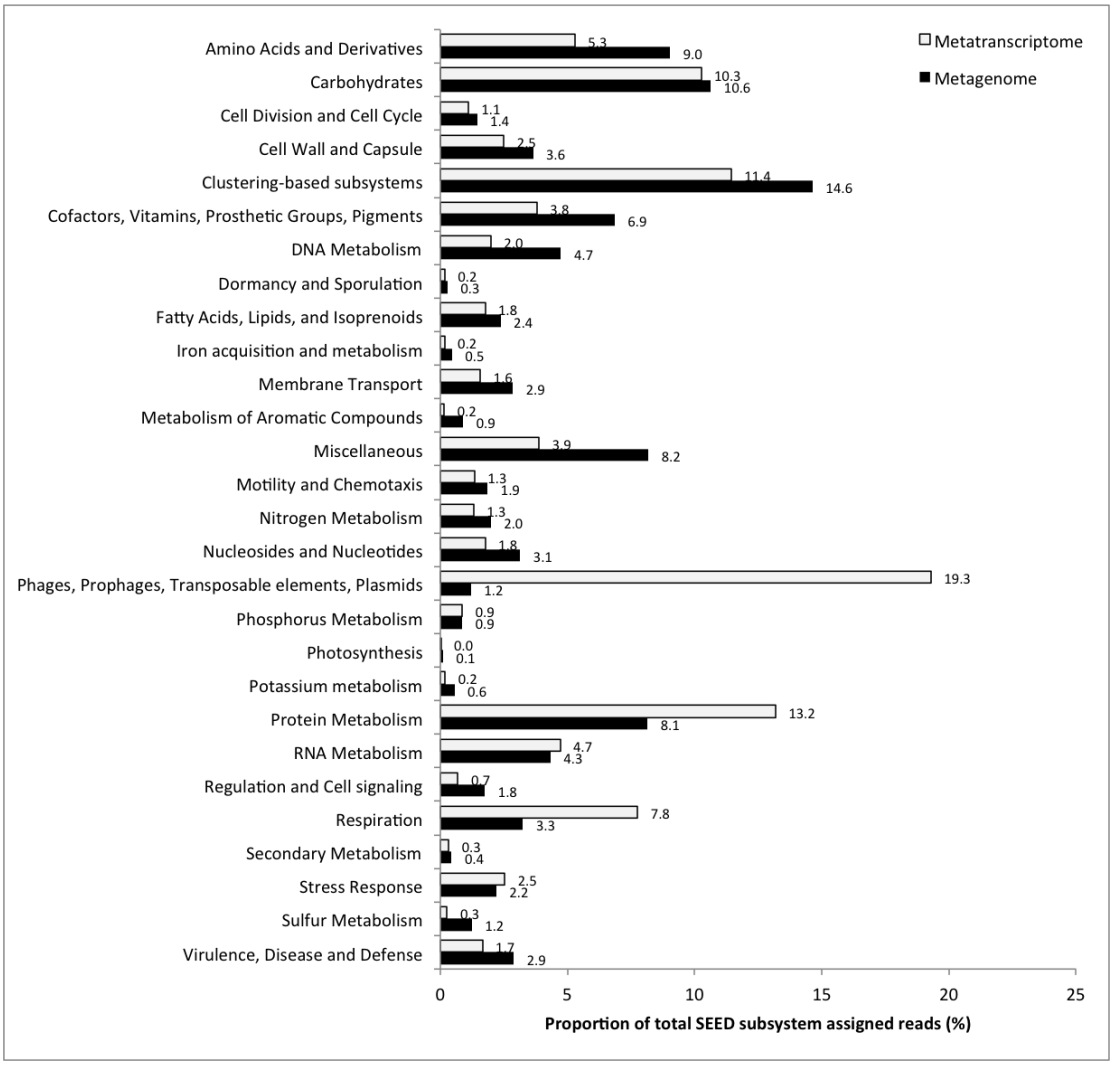


Fig 6.3 Assignments of sequences to SEED subsystems for metatranscriptome and metagenome datasets by MG-RAST

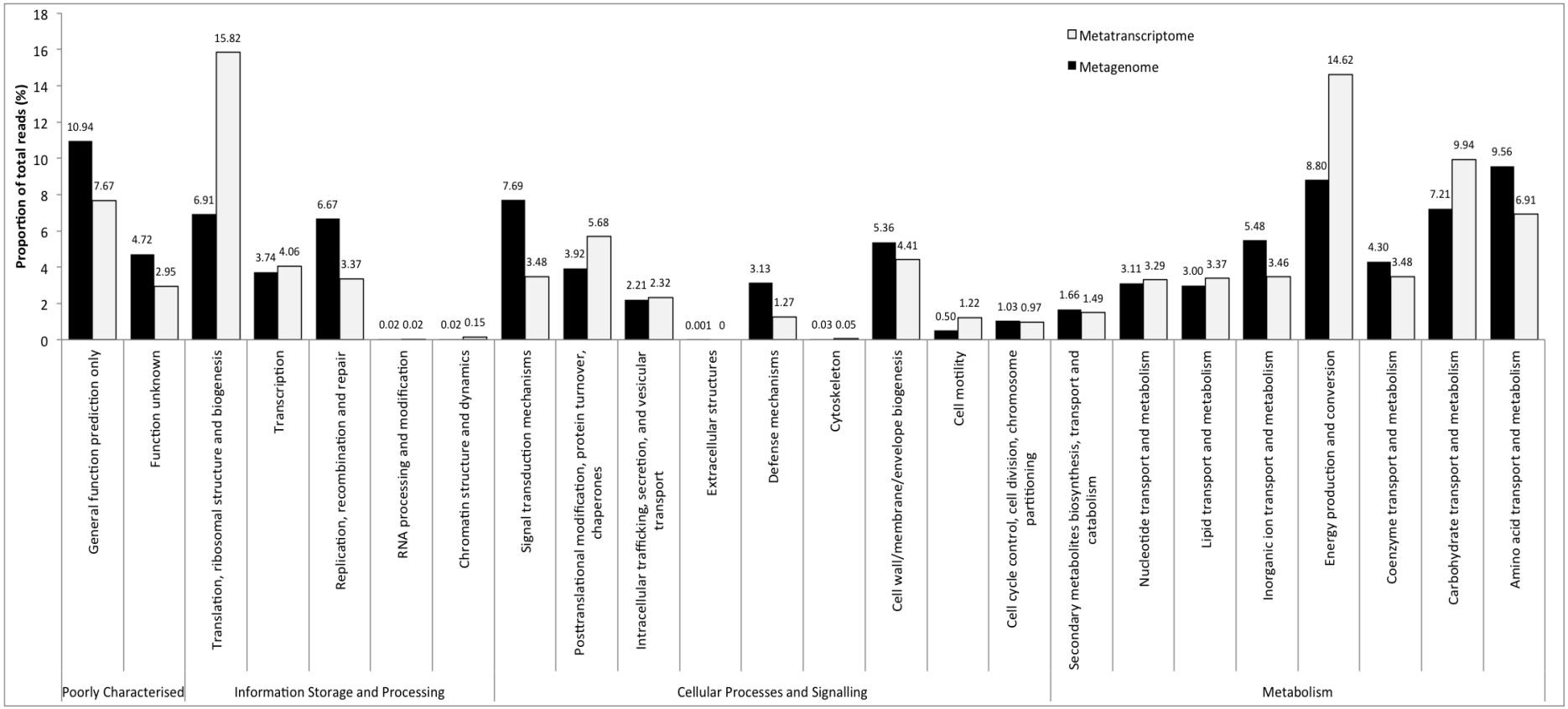


Fig 6.4 Assignment of sequences to COG categories for the metatranscriptome and metagenome datasets by MG-RAST

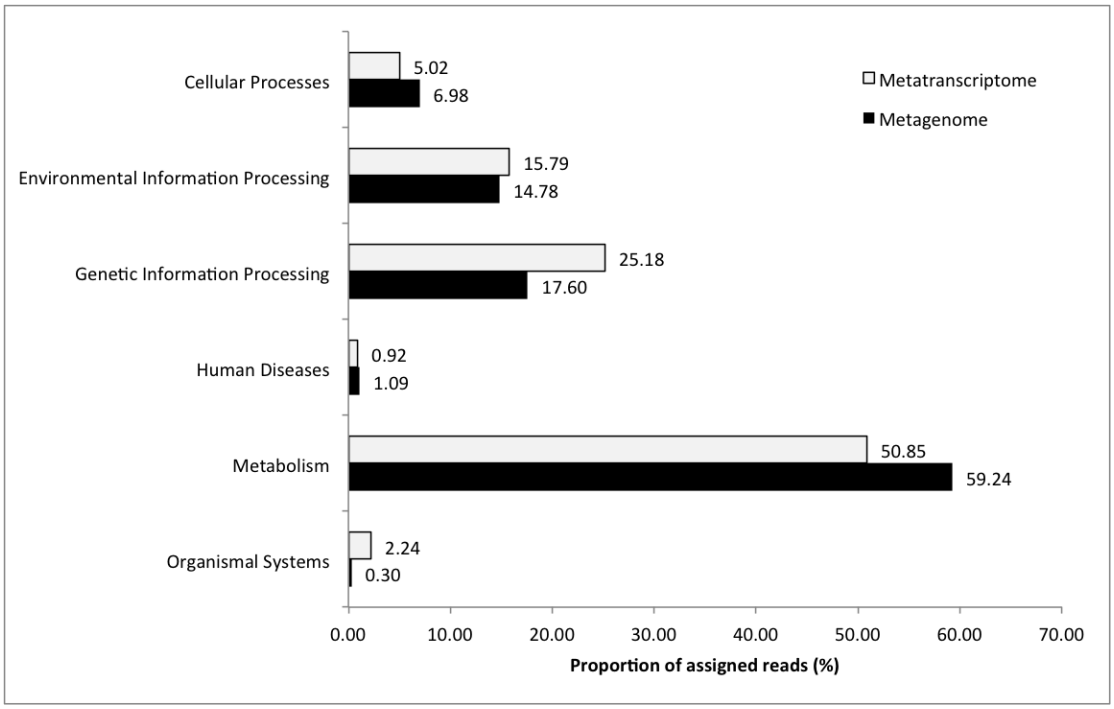


Fig 6.5 Assignment of sequences to KEGG Orthologies for the metatranscriptome and metagenome datasets

6.5.4 KEGG mapping for pathway analysis.

Using the KEGG mapping function of MG-RAST, it was possible to investigate the presence of enzyme functions needed for cellulose utilisation by the microbial community. Fig 6.6 illustrates the KEGG map of starch and sucrose metabolism which includes the pathway by which cellulose is converted to cellobiose units that are in turn broken down to glucose.

Enzyme functions EC 3.2.1.4 and EC 3.2.1.91 both appear in the metagenome and metranscriptome datasets, representing a presence and expression of enzymes required for hydrolysis of 1,4-beta-D-glucosidic linkages found in cellulose, lichenin and also cereal beta-D-glucans. EC 3.2.1.21 is also present, which represents the group of hydrolases responsible for breaking down cellobiose units into beta-D-glucose molecules. As the full pathway for the conversion of cellulose to glucose monomers is present, it would appear that organisms in the community are indeed utilising cellulose.

Interestingly, cellulase synthase activity (EC 2.4.1.12) appeared to be present in the metagenome, but was not expressed in the metratranscriptome.

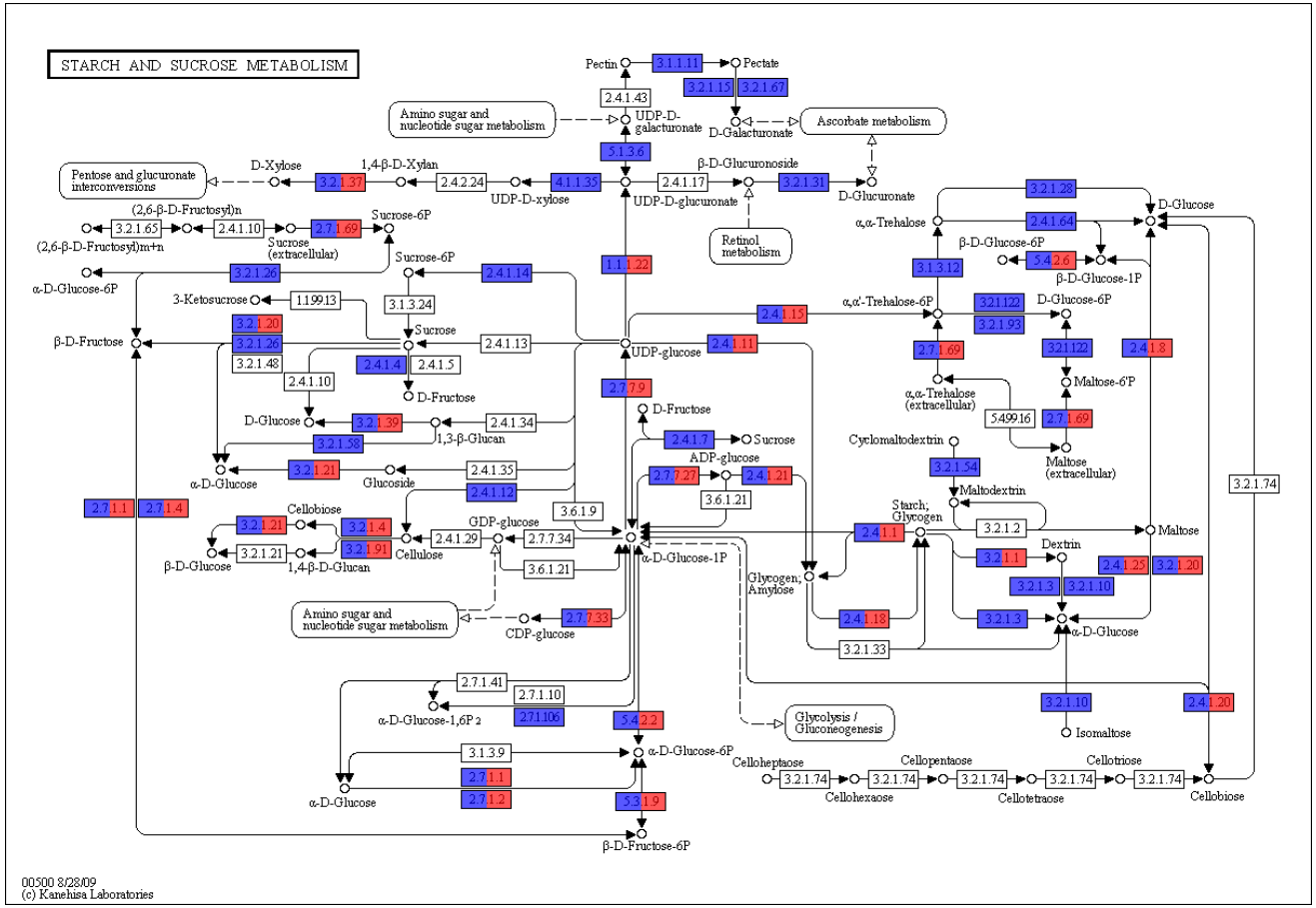


Figure 6.6 KEGG map of the Starch and Sucrose Metabolism pathways. Enzyme categories highlighted in blue were present in the metagenome and red in the metatranscriptome.

6.5.5 Searching the metagenome for expressed genes from the metatranscriptome dataset

ORF prediction with the metatranscriptome reads and sensitive searching against the pfam database had determined 503 sequences with strong homology to glycoside hydrolase families. These sequences are good candidates for representing carbohydrate-active enzymes in the metatranscriptome and Blast results suggested that many sequences were similar to known proteins, but different enough to be potentially distinct, and in some cases the sequences were not found to have a match to the database. As the metatranscriptome and metagenome datasets were produced from the sequencing of material derived from similar samples, from the same geographic site, the samples were produced a year apart and the community samples would not be identical. However, expressed genes in the metatranscriptome indicate good candidates to search for in the metagenomic dataset, where the longer sequences might provide better information concerning the sequence identity and help to design primers to use for the screening of a fosmid library.

The predicted nucleotide sequences of the 503 sequences were used to build a blast database. The metagenome dataset was compared to this database using a blastn search, and hits with a 100% alignment identity were extracted from the output file and summarised. The resulting summary is presented in Table 6.3 and includes the annotation of the highest-scoring blast hit for the original read detected in the metatranscriptome. A total of 112 out of the 503 sequences were found to have exact matches in the metagenome dataset, and some of these had multiple matches. For some of the genes which were well represented in the metagenome, the original metatranscriptome read had been assigned a hypothetical function or a general function only and in one case several reads in the metagenome were apparently identical to a metatranscriptomic read with no close homologue in the database. These reads in particular probably represent genes that merit further study as they are probably carbohydrate active enzymes and might exhibit novel glycoside hydrolase activity either against cellulose or other recalcitrant polysaccharides such as xylan or chitin.

Table 6.3 Metagenomic reads with identical matches to highly probable carbohydrate active enzyme encoding sequences from the metatranscriptome dataset (Table 5.5). ID column refers to the ID assigned to the predicted reads by the metagenemark ORF finder software with the first predicted read being assigned gene_id_1, and so forth. Count is the number of times a metagenomic read had a match at an identity of 100% and the annotation is that originally determined for the metatranscriptomic sequence.

ID*	Count	Genome or Source Organism	Annotation
gene_id_74386	80	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_68027	28	No hits found	
gene_id_25744	26	Dysgonomonas gadei ATCC BAA-286	Hypothetical protein
gene_id_69031	21	Ktedonobacter racemifer DSM 44963	4-alpha-glucanotransferase
gene_id_63225	14	Streptobacillus moniliformis DSM 12112	4-alpha-glucanotransferase
gene_id_50302	12	Roseiflexus castenholzii DSM 13941	4-alpha-glucanotransferase
gene_id_51002	10	Clostridium thermocellum DSM 2360	Beta-glucosidase
gene_id_44728	9	Treponema pallidum subsp. pertenue str. SamoaD	glycosyl hydrolase domain protein
gene_id_43451	8	Geobacillus sp. 70PC53	CelA precursor
gene_id_26698	7	Clostridium novyi NT	4-alpha-glucanotransferase
gene_id_50136	6	Uncultured bacterium	cellodextrinase
gene_id_60574	6	Mucilaginibacter paludis DSM 18603	maltose phosphorylase
gene_id_39528	5	uncultured organism	putative carbohydrate-active enzyme
gene_id_44310	5	beta-1,4-endoglucanase	Pratylenchus vulnus
gene_id_66201	5	Chloroflexus aggregans DSM 9485	glycoside hydrolase family protein
gene_id_11341	4	Clostridium sp. BNL1100	glycosyl hydrolase family 11, dockerin-like protein
gene_id_37232	4	uncultured organism	putative carbohydrate-active enzyme
gene_id_44895	4	Fibrobacter succinogenes subsp. succinogenes S85	glycoside hydrolase family protein
gene_id_66675	4	Acetivibrio cellulolyticus CD2	glycoside hydrolase family protein
gene_id_67840	4	Solitalea canadensis DSM 3403	Endoglucanase
gene_id_7329	4	Alistipes sp. HGB5	conserved hypothetical protein

gene_id_751	4	Mycobacterium kansasii ATCC 12478	Beta-glucosidase
gene_id_11483	3	uncultured bacterium	putative cellulase
gene_id_14125	3	Bacillus cereus E33L	Chitosanase
gene_id_16194	3	Bacteroides uniformis	hypothetical protein
gene_id_2671	3	Treponema brennaborense DSM 12168	Beta-glucosidase
gene_id_31811	3	Anaerophaga thermohalophila DSM 12881	Beta-glucosidase
gene_id_32885	3	Subdoligranulum sp	hypothetical protein
gene_id_37226	3	uncultured bacterium	endo-1,4-beta-xylanase precursor
gene_id_40408	3	Fibrobacter succinogenes subsp. succinogenes S85	Cellulase
gene_id_41114	3	uncultured organism	putative carbohydrate-active enzyme
gene_id_44888	3	uncultured organism	putative carbohydrate-active enzyme
gene_id_53994	3	Cytophaga hutchinsonii ATC 33406	B-glycosidase
gene_id_5816	3	Clostridium hathewayi DSM 13479	4-alpha-glucanotransferase
gene_id_6403	3	Fibrella aestuarina BUZ 2	glycoside hydrolase family 65 central catalytic
gene_id_70631	3	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_14819	2	Dyadobacter fermentans DSM 18053	glycoside hydrolase
gene_id_17156	2	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_2297	2	Butyrivibrio fibrisolvens	Arabinogalactan endo-1,4-beta-galactosidase
gene_id_24494	2	Cytophaga hutchinsonii ATCC 33406	beta-xylosidase/alpha-L-arabinofuranosidase-like protein
gene_id_25713	2	Cyclobacterium marinum DSM 745	glycoside hydrolase
gene_id_29276	2	Trichodesmium erythraeum IMS101	Alpha-glucosidase
gene_id_32572	2	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_34698	2	Paenibacillus curdlanolyticus YK9	glycoside hydrolase family 3 domain protein
gene_id_36845	2	Roseburia intestinalis XB6B4	Alpha-glucosidase
gene_id_40794	2	Bacillus circulans	Beta-glucanase
gene_id_4168	2	Fibrobacter succinogenes subsp. succinogenes S85	glycoside hydrolase family protein
gene_id_48353	2	Alistipes sp. JC136	maltose phosphorylase
gene_id_60617	2	Melioribacter roseus P3M	beta-glucanase precursor

gene_id_61935	2	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_64127	2	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_64445	2	Dysgonomonas gadei ATCC BAA-286	hypothetical protein
gene_id_65839	2	Bacteroides salanitronis DSM 18170	alpha-1,2-mannosidase
gene_id_68318	2	Flavobacterium sp.	Por secretion system C-terminal sorting domain-containing protein
gene_id_69147	2	Haliangium ochraceum DSM 14365	glycoside hydrolase
gene_id_71795	2	Ignavibacterium album JCM 16511	beta-glucosidase
gene_id_78222	2	uncultured bacterium	glycoside hydrolase family 5 protein
gene_id_8996	2	Anaerophaga thermohalophila DSM 12881	beta-glucosidase
gene_id_1013	1	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_11643	1	Candidatus Solibacter usitatus Ellin6076	glycoside hydrolase family 3 protein
gene_id_11669	1	Cytophaga hutchinsonii ATCC 33406	Xylanase
gene_id_1225	1	Saccharophagus sp. Myt-1	Cellulase
gene_id_15282	1	Trichodesmium erythraeum IMS101	Alpha-glucosidase
gene_id_1548	1	Sulfolobus islandicus L.D.8.5	glycoside hydrolase family protein
gene_id_17281	1	Cellvibrio sp. BR	endo-beta-1,3(4)-glucanase
gene_id_18101	1	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_18246	1	Ruminococcus sp. SR1/5	4-alpha-glucanotransferase
gene_id_18937	1	uncultured bacterium	hypothetical protein
gene_id_19005	1	Spirochaeta caldaria DSM 7334	glycoside hydrolase family protein
gene_id_2092	1	Cytophaga hutchinsonii ATCC 33406	B-glycosidase
gene_id_21988	1	Clostridium papyrosolvens DSM 2782	glycoside hydrolase family 8 protein
gene_id_23067	1	Flavobacterium sp.	Por secretion system C-terminal sorting domain-containing protein
gene_id_23904	1	Cytophaga hutchinsonii ATCC 33406	beta-glycosidase-like protein
gene_id_24207	1	Spirochaeta caldaria DSM 7334	Cellulase
gene_id_26637	1	Mucilaginibacter paludis DSM 18603	arabinogalactan endo-1,4-beta-galactosidase
gene_id_30687	1	Cytophaga hutchinsonii ATCC 33406	B-glycosidase
gene_id_32467	1	Trehalose/maltose hydrolase (phosphorylase)	Flavobacteriales bacterium ALC-1

gene_id_33242	1	uncultured organism	putative carbohydrate-active enzyme
gene_id_34778	1	Clostridium clariflavum DSM 19732	beta-glucosidase-like glycosyl hydrolase
gene_id_39547	1	uncultured organism	putative carbohydrate-active enzyme
gene_id_42287	1	Prevotella ruminicola 23	family 25 glycosyl hydrolase
gene_id_42923	1	Niastella koreensis GR20-10	4-alpha-glucanotransferase
gene_id_44553	1	Herpetosiphon aurantiacus DSM 785	glycoside hydrolase
gene_id_488	1	Arabidopsis lyrata subsp. lyrata	glycoside hydrolase family 3 protein
gene_id_49758	1	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_50341	1	Clostridium sp. BNL1100	beta-xylosidase
gene_id_50487	1	Treponema brennaborense DSM 12168	4-alpha-glucanotransferase
gene_id_51842	1	uncultured organism	putative carbohydrate-active enzyme
gene_id_52371	1	Flavobacteriaceae bacterium S85	glycoside hydrolase family protein
gene_id_52939	1	Prevotella ruminicola 23	family 25 glycosyl hydrolase
gene_id_52980	1	Turicibacter sp. HGF1	putative chitinase ChiB1
gene_id_55301	1	Cytophaga hutchinsonii ATCC 33406	beta-xylosidase/alpha-L-arabinofuranosidase-like protein
gene_id_57755	1	Anaerophaga sp. HS1	mannan endo-1,4-beta-mannosidase
gene_id_58109	1	Dickeya dadantii Ech703	Cellulase
gene_id_59545	1	Bacteroides helcogenes P 36-108	glycoside hydrolase family 3 protein
gene_id_61108	1	Dysgonomonas gadei ATCC BAA-286	hypothetical protein
gene_id_61989	1	Streptomyces sp. W007	glycosyl hydrolase
gene_id_63263	1	Caldicellulosiruptor owensensis OL	xylan 1,4-beta-xylosidase
gene_id_64798	1	Aquimarina agarilytica ZC1	Cellulase
gene_id_66031	1	Physcomitrella patens subsp. Patens	predicted protein
gene_id_6662	1	Clostridium josui	Endoglucanase
gene_id_72064	1	Cytophaga hutchinsonii ATCC 33406	beta-glycosidase-like protein
gene_id_72411	1	Eubacterium siraeum	Beta-mannanase
gene_id_73258	1	uncultured organism	putative carbohydrate-active enzyme
gene_id_73835	1	Cytophaga hutchinsonii ATCC 33406	bifunctional acetylxylan esterase/xylanase
gene_id_7666	1	Anaerophaga thermohalophila DSM 12881	glycoside hydrolase family protein

gene_id_7682	1	uncultured bacterium	glycoside hydrolase family 3protein
gene_id_77038	1	Cytophaga hutchinsonii ATCC 33406	bifunctional acetylxylan esterase/xylanase
gene_id_77460	1	Anaerophaga thermohalophila DSM 12881	glycoside hydrolase family protein
gene_id_7860	1	Cytophaga hutchinsonii ATCC 33406	endoglucanase-like protein
gene_id_85521	1	Eubacterium siraeum	hypothetical protein
gene_id_86368	1	Paludibacter propionicigenes WB4	glycoside hydrolase

6.6 Discussion

A metagenome was extracted from colonised cotton baits and sequenced as the metatranscriptome had been previously. The metagenomic sequencing was carried out after the metatranscriptome had already been completed and on material produced by a later sampling campaign which means that a direct comparison between the two should be made with some level of caution. Nevertheless, using metatranscriptomic and metagenomic sequence on similar samples from the same environment should shed some light on the inherent biology of the cellulose colonising biofilm.

Upgrades to the instrumentation available at the CGR, and the implementation of a basic level of quality screening performed on their sequence output files, also slightly altered the way the data files were handled. Sequencing technology development moves at a fast pace however and increases to sequencing data output have to be taken in stride. The sequencing undertaken here, to gain an overview of the biology of the cellulose bait colonising bacteria and screen the data for potentially interesting carbohydrate active enzymes with roles in the breakdown of cellulose does not really depend on the starting material being processed in an identical manner. In a sense, the analysis is much more qualitative than quantitative.

Ongoing updates of software and databases means that analyses performed a few months apart will be different and this can also cause small amounts of variation. Impossible to keep returning to old datasets to constantly re-analyse them but careful design should to a great extent alleviate this issue, and ensure datasets are produced and analysed as concurrently as possible. This is another issue with the use of a webserver for analysis; updates to programs and databases will proceed automatically but locally installed software and databases can be left in their currant state and updates avoided for a time for the sake of treating related datasets the same way.

In order to reduce long upload time for transferring data to the MG-RAST webserver, quality information was removed from the data files before submission (i.e. the data was converted from fastq format to fasta format). Prinseq can output data as fastq and/or fasta format and after QC of the metagenome fasta format data was produced, which shrinks the size of the file as quality information had been removed and renders uploading the file to the MG-RAST webserver a faster process. Although the removal of quality score information limits the quality-control options

that can then be implemented by the MG-RAST pipeline, this is not really an issue as the dataset had been extensively quality-filtered.

As the metatranscriptome and metagenome were not, unfortunately, produced concurrently and from the same samples differences between the two datasets in many instances may not represent difference between expression levels and gene content. Randomness and stocastity of colonisation of the cellulose baits by environmental microbes might explain some of the variation seen. Natural communities of microorganisms are constantly experiencing changing conditions and numbers reads of specific species and specific gene transcripts are subject to temperal variation (Andersson *et al*., 2010; Gilbert *et al*., 2009). Some differences could be caused by the sampling process, which involved dropping weighted baits randomly into the deeper part of the lake. Some baits might end up deep in the sediment layers while others might be situated nearer the sediment-water interface.

There was an apparently great amount of viral activity detected in the metatranscriptome, much of which stemming from active gene expression of phages related to Enterobacteria Phage PhiX174. This was not mirrored in the analysis of the metagenome. Some viruses do have RNA genomes and would not show up in the metagenome dataset, but would be present in the metatranscriptome, as a result of the sampling process. In this vein, some sequences classified as viral in the metatranscriptome might in fact constitute RNA virus genomic material and not be a signal of active ongoing infections, although the majority of phage described from aquatic environments have been DNA viruses (Weinbauer, 2004). One source of discrepancy in numbers of viral sequences might be related to the detection of ongoing infections where viruses have hijacked host machinery to transcribe and translate their genes at high efficiency and produce thousands of progeny so that many more copies of transcripts that genes are present; it makes sense that viruses should be detected at a greater level in the metatranscriptome dataset. The family Microviridae to which the Enterobacteria Phage belong reportedly normally replicate using the lytic cycle rather than the lysogenic as temperate phage from this family are very much in the minority (Roux *et al*. 2012).

Although the datasets were produced from separate sampling campaigns, performed a year apart from one another, the patterns of prevalence of bacteria at phylum level are broadly similar. The three most abundant phyla were the same across the two datatsets and there was generally no more than a two or threefold difference in the abundance of any of the phyla between the datasets. There was a much greater amount of variation in the abudances of Eukaryotic phyla. The huge spike in metatranscriptomic reads of the phylum Platyhelmininthes, and the almost

complete lack of any such sequences in the metagenome suggests a temporary local abundance of flatworms at the time of sampling. Many of the Eukaryotic sequences appeared to represent a background noise of normal residents of the lake ecosystem such as fish and algae, which were for the most part probably not specifically associated with the cotton baits.

More importantly, sequences from the metatranscriptome representing probable glycoside hydrolase sequences were confirmed to be present in the metagenome. That identical sequences should be retrieved from both of these independent datasets is compelling evidence that the genes that these sequences represent are numerous in the environment and the species that have them are probably of some functional importance. As both the metatranscriptome and metagenome were derived from microorganisms in association with a cellulose substrate, this increases the likelihood that these sequences represent gene with a genuine role in cellulose degradation.

The fosmid library was produced in chapter 4 was generated from DNA extracted from the same material as the metagenome sequenced here. It is therefore likely that some of the glycoside hydrolase sequences detected *in silico* might be present in some of these fosmids. Expression screening of the fosmids might lead to the identification of carbohydrate active genes, as has been achieved previously with fosmid clones (e.g. Geng *et al*., 2012). Use of the sequence information from the metatranscriptome and metagenome can also be used to design primers to perform PCR screening of the fosmid library to avoid the potential issue of unexpressed genes in fosmid clones or specific activities remaining undetected by screening methods.

Most studies employ either metagenomic or metatranscriptomic sequencing but not both. In the studies which have done both, the picture that emerges is that phylogenetic groups that are found to be abundant though metagenomic sequencing will also be found in abudance in the associated metatranscriptome, although the levels of abundance will be different between the two (Yu & Zhang, 2012; Shi *et al*., 2011). The results presented here appear to agree with previous published observations in that regard.

Chapter 7: General Discussion

Quantitative PCR and high-throughput sequencing of a metagenome and a metatranscriptome have been used to study communities of microorganisms in anaerobic environments. The work has specifically focussed on cellulolytic organisms and has sought to assess the presence of microorganisms responsible for breaking down cellulose in landfill leachate and the sediment of a freshwater lake and to identify the genes expressed by these organisms which are involved in enzymatic digestion of cellulose polymers. Some of the main findings of the study are as follows:

- Bacteria that are members of or closely related to the genus *Fibrobacter* have been demonstrated to be present in landfill leachate and lake sediment, indicating a wider distribution throughout the environment than previously appreciated.
- *Fibrobacter spp.* have additionally been demonstrated to have an important role in cellulose breakdown in landfill as the organism was enriched in the biofilm colonising a cotton cellulose baits which was rapidly and efficiently degraded and either absent or present only in very low numbers in the biofilm colonising a cellulose bait which was not at all degraded by the colonising microorganisms. This demonstrates that *Fibrobacter* not only exists outside of the gut environment but has a similar cellulolytic role in communities in other environments.
- The marked different in colonisation between the two cotton cellulose baits in separate microcosms indicates that major difference in microbial populations can exist even between similar environments.
- Metatranscriptomic sequencing of the microbial community associated with cellulose baits lake sediment revealed a community dominated primarily by bacteria and identified an active presence of cellulolytic lineages including *Fibrobacter spp.*, *Clostridium spp.*, *Ruminococcus spp.* and *Cytophaga hutchinsonii*.
- Metatranscriptomic sequencing also identified bacteriophage as highly active members of the ecosystem.
- Glycoside hydrolase sequences were detected in the lake sediment metatranscriptome by homology to protein families and many of these had only partial matches to entries in the NCBI nr database suggesting that they

exhibited a degree of novelty. Several sequences were not found to have representatives in the database at all.

- Comparative analysis revealed that the metatranscriptome and metagenome was both dominated by bacteria, but Eukaryotic and viral sequences were less well represented by the metagenome. At the bacterial phylum level, the two datasets were broadly similar.
- Comparative analysis also revealed of a metagenome and a metatranscriptome revealed that *Fibrobacter spp.* and *Cytophaga hutchinsonii* were better represented in the metatranscriptome than the metagenome and that the reverse was true for the clostridia, which suggests the former are the key cellulolytic species in this environment. *Fibrobacter* and *C. hutchinsonii* have been hypothesised to possess as yet uncharacterised mechanisms for cellulose degradation and their active presence associated with the cellulose baits from the lake and the detection of glycoside hydrolase sequences with little or no similarity to those already known suggests that novel enzymatic mechanisms might be active in this environment.

This study has added to the understanding of the microbial degradation of cellulose in anaerobic environments outside of the gut. Previously cellulose breakdown in landfill had been thought to have been mediated mostly by members of the clostridia (Burrell *et al*. 2004, O'Sullivan *et al*. 2007). It was once thought that, outside of the rumen, anaerobic degradation of cellulose was mainly mediated by clostridia and some members of the Bacteroidetes (Leschine *et al*. 1995) but evidence is mounting that the bacterial phylum Fibrobacteres and fungal phylum Neocallimastigales are not limited to the rumen (Ransom-Jones *et al*. 2012, Van Dyke & McCarthy 2002), and environmental members of these groups might be found more widely, if sought. The evidence presented here certainly supports a wider distribution of *Fibrobacter spp*. Additionally the functional glycoside hydrolase genes identified in the data show only limited similarity to entries in the NCBI NR database which could indicate a level of novelty if not in functional terms then in terms of traits such as efficient enzymatic activity at low temperatures.

The cellulolytic bacteria of landfill were originally studied as a consequence of their role in production of methane gas (Westlake *et al*. 2005). Methane is produced in landfill sites as the result of microbial activity and begins with the decomposition of organic material (Barlaz, 2006). Typically the most abundant organic material in landfill is cellulosic and hemicellulosic material. The harvesting of methane gas from landfill sites can been explored as a method of reclaiming energy (Zamorano *et al*.,

2007). The study of cellulolytic bacteria and the attempt to discover new cellulases has received a great deal of attention lately due to a growing requirement for renewable energy. Biofuels have been developed as a replenishable alternative to fossil fuels. The first generation of biofuels were produced by fermentation of food crops which is undesirable as it affects food prices (Hill *et al*., 2006). Cellulosic biofuels can be generated from the microbial breakdown of cellulosic material such as agricultural waste or fuel crops such as switchgrass which can be grown on marginal land and therefore not compete with the human supply chain. Cellulosic biofuel, though, has not yet proven economical to produce mainly because the process of enzymatically breaking down cellulase polymers into fermentable sugars that can in turn be converted to useful fuel such as ethanol is expensive and inefficient. Alternative methods such as consolidated bioprocessing are being developed where a single organism, or a consortium of organisms, are used to breakdown cellulose and convert the breakdown products in fuel in a single step (Higashide *et al*., 2011; Lynd *et al*., 2005). Studying cellulose degrading communities which have evolved naturally could lead to the discovery of new organisms that can be used in this process or identify new genes that could be engineered into strains developed for use in industrial-scale consolidated bioprocessing.

The analysis of the datasets performed at present is limited to the identification of sequences *in silico* with statistical likelihoods of belonging to specific Pfam families, or being similar to a particular entry in the NCBI nr database. Additionally, a fosmid library has been produced but in no way screened for the presence of potentially interesting enzymes. Expression screening has been used by many metagenomic studies to search for cellulases in fosmid libraries and can certainly be employed here as it has been used to identity other carbohydrate active enzymes present in metagenomic fosmid libraries (Rashamuse et al. 2013). However, it is conceivable that such an approach could miss some interesting biology as enzymes might require other genes to be present, or need specific pre-processing and secretion pathways not necessarily present in the fosmid host strain, to exhibit their function. The sequence information from the metagenome and metatranscriptome can also be used to design primers for a molecular interrogation of the gene content of the fosmid library, and may lead to the identification of fosmids containing full-length genes represented by some of the sequence reads.

Many previous metagenomic efforts to search for new cellulases have employed production of fosmid libraries and screening these libraries for activity and other studies have utilised high throughput sequencing of metagenomic DNA but only very few examples of the use of metatranscriptomics for the discovery of novel

cellulases exist. The studies have been done have focussed on aerobic fungal cellulose degraders in soils and leaf litter (Zifcakova & Baldrian, 2012). This use of metatranscriptomics, in addition to fosmid library production and metagenomics, is the first time metatranscriptomic sequencing has been used as a method to specifically search for cellulases in an environment where much of the cellulose is broken down by bacteria, not fungi.

It is perhaps understandable that metatranscriptomics has not yet been used in this capacity as it is a much more technically challenging approach. Bacterial mRNA transcripts lack the polyA tails of eukaryotes. Approaches such as the capture of mRNA can be captured with bound polydT sequences as employed by Takasaki *et al*. (2013) cannot be applied and the specific selection of mRNA bacterial communities impossible and large amounts of unwanted rRNA must be contended with. However, it is possible that metatranscriptomic sequencing might become more common in microbial ecology projects as many recent developments have greatly facilitated the process of sequence library generation. Although efficient rRNA removal has been a traditional stumbling block for environmental metatranscriptomic studies focused on bacterial communities, Ribo-Zero technology from Epicentre appears to be able to reproducibly lower rRNA content in metatranscriptomic sequencing analysis to less than 5% of the total reads and to as low as 1% if the RNA is of good integrity (Giannoukos *et al*. 2012). The Ribo-Zero rRNA depletion kit, unavailable when the metatranscriptome sequenced here was first produced, would likely have done a better job of rRNA removal than the Terminator exonuclease used to digest rRNA sequences and will certainly be extensively used in the coming years to remove excess rRNA from total RNA preparations and ensure mRNA sequences are well represented in metatranscriptome sequencing.

In addition to improved rRNA removal, the necessity of obtaining large amounts of high quality RNA for sequencing is now less problematic. In 2011 when work to produce the metatranscriptome was begun the CGR required a dscDNA sample consisting of 1.5 – 3µg of material for sequencing library preparation. Multiple extractions of RNA were pooled together, but after cleanup and DNAse treatment only approximately 1 µg of RNA was available for further applications and amplification of this RNA using the MessageAmp kit was necessary to generate large amounts of amplified RNA which could be used in turn to produce cDNA libraries. Within the past twelve months the ScriptSeq library preparation kits have become available, which can be used to produce a cDNA library for sequencing on the Illumina platform and the CGR will accept an rRNA-depleted or polyadenylated sample of RNA as small as 10 ng. The ScriptSeq manufacturers specify that using

this kit a sequencing library can be produced using as little as 0.5 ng of starting material although 50 ng is preferred. The use of this kit removes the need to amplify the RNA sample, convert amplified RNA to cDNA and then produce a sequencing library and as such cuts down on time and sample handling.

The use of metagenomic and metatranscriptomic sequencing for gene discovery is particularly advantageous over other molecular techniques. PCR requires the design of specific primers and as such can be highly specifically targeted to specific groups or functional genes and here primer specificity was used to detect several separate groups of bacteria using qPCR. Primer design must always be based on currently known sequences present in the databases, a limitation which means detection of unknown truly novel sequences is impossible using PCR-based analyses in microbial ecology (Smith & Osbourne, 2009). Cloning can identify novelty but thousands if not tens of thousands of clones are produced in such studies and generally yield only one or two interesting enzymes (Rashamuse *et al*., 2013; Geng *et al*., 2012; Liu *et al*., 2011). High-throughput sequencing can identify hundreds of potential cellulase sequences (Xia *et al*., 2013; Hess *et al*., 2011) although cloning does retain the advantage that full length biochemically characterised genes are obtained rather than just data.

Metatranscriptomic analysis of cellulose degrading fungal communities has detected glycoside hydrolase gene reads at rate of 0.5-0.8% of the total (Zifcakova & Baldrian, 2012), and cloning studies might identify one CAZy gene in every thousand or ten thousand clones. The metatranscriptomic sequencing of cotton baits in a lake sediment here identified 503 glycoside hydrolases out of just under one million sequences, a hit rate of 0.0005%. The sheer number of reads generated by Illumina sequencing means that only a tiny fraction of the reads need correspond to sequences of interest to produce interesting data which is certainly a strength. Longer incubation times for the cellulose baits might have increased the proportion still further, however.

It is possible that eventually the types of metagenomic and metatranscriptomic sequencing studies currently published will be largely replaced by single-cell sequencing experiments. Transcriptome sequencing of the mRNA of a single cell has been feasible for a number of years now and, importantly, generation of sufficient quantities of cDNA for sequencing applications can be achieved without introducing bias even when working with the mRNA content of only one cell (Tang *et al*. 2009). Recently Rinke *et al*. (2013) reported the use of single-cell genomics to study uncultivatable cells from diverse environments. The use of this technique to investigate microbial cells in the environment has advantages over metagenomic

studies relying on high-throughput sequencing; studying the genome or transcriptome of a single cell produces a much more detailed and thorough analysis of the genes and gene expression therein, as the issues of coverage which plague metagenomics and metatranscriptomics can be avoided. The dataset produced from the genome or transcriptome of a single will also be much simpler to analyse. It is known all of the reads originate from one organism and assembly is much more feasible, and blast analysis of the data is simpler; a match to a specific database entry might be due to a sequence that has no good representative in the database being aligned with distant relative as a closest match and analysis of reads from a single cell will give much greater context to the interpretation of such results. This type of experiment may become more useful than sequencing a metagenome to obtain a partial, fragmentary view of a complex community. Ultimately, it is developments in sequencing technology that will probably determined the best way to understand the ecology of microbes in their natural habitats. In 2000 nobody anticipated the routine production of datasets containing millions of sequences in the form of gigabytes of data. In another ten years it is likely that the available technology and methodology for microbial ecology will once again have evolved.

References

Alonso, D. M., Bond, J. Q. & Dumesic, J. A. (2010). Catalytic conversion of biomass to biofuels. *Green Chemistry* **12**, 1493-1513.

Amend, A. S., Seifert, K. A. & Bruns, T. D. (2010). Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology* **19**, 5555-5565.

Andersson, A. F., Riemann, L. & Bertilsson, S. (2010). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *Isme Journal* **4**, 171-181.

Barlaz, M. A. (2006). Forest products decomposition in municipal solid waste landfills. *Waste Management* **26**, 321-333.

Beguin, P. & Aubert, J. P. (1994). The Biological Degradation of Cellulose. *Fems Microbiology Reviews* **13**, 25-58.

Bomar, L., Maltz, M., Colston, S. & Graf, J. (2011). Directed Culturing of Microorganisms Using Metatranscriptomics. *Mbio* **2**,

Bomble, Y. J., Beckham, G. T., Matthews, J. F., Nimlos, M. R., Himmel, M. E. & Crowley, M. F. (2011). Modeling the Self-assembly of the Cellulosome Enzyme Complex. *Journal of Biological Chemistry* **286**, 5614-5623.

Bonot, S., Courtois, S., Block, J. C. & Merlin, C. (2010). Improving the recovery of qPCR-grade DNA from sludge and sediment. *Applied Microbiology and Biotechnology* **87**, 2303-2311.

Brulc, J. M., Antonopoulos, D. A., Miller, M. E. B., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., Edwards, R. E., Frank, E. D., Emerson, J. B., Wacklin, P., Coutinho, P. M., Henrissat, B., Nelson, K. E. & White, B. A. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 1948-1953.

Burrell, P. C., O'Sullivan, C., Song, H., Clarke, W. P. & Blackall, L. L. (2004). Identification, detection, and spatial resolution of Clostridium populations responsible for cellulose degradation in a methanogenic landfill leachate bioreactor. *Applied and Environmental Microbiology* **70**, 2414-2419.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research* **37**, D233-D238.

Cheng, H. R. & Jiang, N. (2006). Extremely rapid extraction of DNA from bacteria and yeasts. *Biotechnology Letters* **28**, 55-59.

Collins, M. D., Lawson, P. A., Willems, A., Cordoba, J. J., Fernandezgarayzabal, J., Garcia, P., Cai, J., Hippe, H. & Farrow, J. A. E. (1994). The Phylogeny of the genus *Clostridium* - Proposal of 5 new genera and 11 new species combinations. *International Journal of Systematic Bacteriology* **44**, 812-826.

Daly, K. & Shirazi-Beechey, S. P. (2003). Design and evaluation of group-specific oligonucleotide probes for quantitative analysis of intestinal ecosystems: their application to assessment of equine colonic microflora. *Fems Microbiology Ecology* **44**, 243-252.

Damon, C., Lehembre, F., Oger-Desfeux, C., Luis, P., Ranger, J., Fraissinet-Tachet, L. & Marmeisse, R. (2012). Metatranscriptomics Reveals the Diversity of Genes Expressed by Eukaryotes in Forest Soils. *Plos One* **7**,

Davenport, C. F., Neugebauer, J., Beckmann, N., Friedrich, B., Kameri, B., Kokott, S., Paetow, M., Siekmann, B., Wieding-Drewes, M., Wienhoefer, M., Wolf, S., Tuemmler, B., Ahlers, V. & Sprengel, F. (2012). Genometa - A Fast and Accurate Classifier for Short Metagenomic Shotgun Reads. *Plos One* **7**,

Dawson, S. C. & Pace, N. R. (2002). Novel kingdom-level eukaryotic diversity in anoxic environments. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 8324-8329.

De Filippo, C., Ramazzotti, M., Fontana, P. & Cavalieri, D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in Bioinformatics* **13**, 696-710.

de Menezes, A. B., McDonald, J. E., Allison, H. E. & McCarthy, A. J. (2012). Importance of Micromonospora spp. as Colonizers of Cellulose in Freshwater Lakes as Demonstrated by Quantitative Reverse Transcriptase PCR of 16S rRNA. *Applied and Environmental Microbiology* **78**, 3495-3499.

de Menezes, A. B., Lockhart, R. J., Cox, M. J., Allison, H. E. & McCarthy, A. J. (2008). Cellulose Degradation by Micromonosporas Recovered from Freshwater Lakes and Classification of These Actinomycetes by DNA Gyrase B Gene Sequencing. *Applied and Environmental Microbiology* **74**, 7080-7084.

Deng, W., Xi, D., Mao, H. & Wanapat, M. (2008). The use of molecular techniques based on ribosomal RNA and DNA for rumen microbial ecosystem studies: a review. *Molecular Biology Reports* **35**, 265-274.

DeSantis, T. Z., Hugenholtz, P., Keller, K., Brodie, E. L., Larsen, N., Piceno, Y. M., Phan, R. & Andersen, G. L. (2006). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research* **34**, W394-W399.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.

Edwards, J. L., Smith, D. L., Connolly, J., McDonald, J. E., Cox, M. J., Joint, I., Edwards, C. & McCarthy, A. J. (2010). Identification of Carbohydrate Metabolism Genes in the Metagenome of a Marine Biofilm Community Shown to Be Dominated by Gammaproteobacteria and Bacteroidetes. *Genes* **1**, 371-384.

Edwards, U., Rogall, T., Blocker, H., Emde, M. & Bottger, E. C. (1989). Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucl. Acids Res.* **17**, 7843-7853.

Engelbrektson, A., Kunin, V., Wrighton, K. C., Zvenigorodsky, N., Chen, F., Ochman, H. & Hugenholtz, P. (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *Isme J* **4**, 642-647.

Ferrer, M., Ghazi, A., Beloqui, A., Vieites, J. M., Lopez-Cortes, N., Marin-Navarro, J., Nechitaylo, T. Y., Guazzaroni, M. E., Polaina, J., Waliczek, A., Chernikova, T. N., Reva, O. N., Golyshina, O. V. & Golyshin, P. N. (2012). Functional Metagenomics Unveils a Multifunctional Glycosyl Hydrolase from the Family 43 Catalysing the Breakdown of Plant Polymers in the Calf Rumen. *Plos One* **7**,

Ferrer, M., Martinez-Abarca, F. & Golyshin, P. N. (2005). Mining genomes and 'metagenomes' for novel catalysts. *Current Opinion in Biotechnology* **16**, 588-593.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. & Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research* **38**, D211-D222.

Franks, A. H., Harmsen, H. J. M., Raangs, G. C., Jansen, G. J., Schut, F. & Welling, G. W. (1998). Variations of bacterial populations in human feces measured by fluorescent in situ hybridization with group-specific 16S rRNA-Targeted oligonucleotide probes. *Applied and Environmental Microbiology* **64**, 3336-3345.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W. & DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 3805-3810.

Fontes, C. & Gilbert, H. J. (2010). Cellulosomes: Highly Efficient Nanomachines Designed to Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annual Review of Biochemistry, Vol 79* **79**, 655-681.

Galand, P. E., Casamayor, E. O., Kirchman, D. L. & Lovejoy, C. (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 22427-22432.

Geng, A., Zou, G., Yan, X., Wang, Q., Zhang, J., Liu, F., Zhu, B. & Zhou, Z. (2012). Expression and characterization of a novel metagenome-derived cellulase Exo2b and its application to improve cellulase activity in Trichoderma reesei. *Applied Microbiology and Biotechnology* **96**, 951-962.

George, D. G., Talling, J. F. & Rigg, E. (2000). Factors influencing the temporal coherence of five lakes in the English Lake District. *Freshwater Biology* **43**, 449-461.

Ghai, R., Martin-Cuadrado, A. B., Molto, A. G., Heredia, I. G., Cabrera, R., Martin, J., Verdu, M., Deschamps, P., Moreira, D., Lopez-Garcia, P., Mira, A. & Rodriguez-Valera, F. (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *Isme Journal* **4**, 1154-1166.
Gifford, S. M., Sharma, S., Rinta-Kanto, J. M. & Moran, M. A. (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *Isme Journal* **5**, 461-472.

Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Huse, S. & Joint, I. (2009). The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology* **11**, 3132-3139.

Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W. Z., Gilna, P. & Joint, I. (2008). Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *Plos One* **3**,

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**, 759-769.

Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *Isme Journal* **3**, 1314-1317.

Gray, N. D., Hastings, R. C., Sheppard, S. K., Loughnane, P., Lloyd, D., McCarthy, A. J. & Head, I. M. (2003). Effects of soil improvement treatments on bacterial community structure and soil processes in an upland grassland soil. *Fems Microbiology Ecology* **46**, 11-22.

Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G. & Bailey, M. J. (2000). Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Applied and Environmental Microbiology* **66**, 5488-5491.

He, Y., Zhao, Y., Zhou, G. & Huang, M. (2009). Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from aged refuse for microbial community analysis. *World Journal of Microbiology & Biotechnology* **25**, 2043-2051.

He, S. M., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., Wang, Z., Chen, F., Lindquist, E. A., Sorek, R. & Hugenholtz, P. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods* **7**, 807-U858.

Higashide, W., Li, Y., Yang, Y. & Liao, J. C. (2011). Metabolic Engineering of Clostridium cellulolyticum for Production of Isobutanol from Cellulose. *Applied and Environmental Microbiology* **77**, 2727-2733.

Hill, J., Nelson, E., Tilman, D., Polasky, S. & Tiffany, D. (2006). Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 11206-11210.

Hong, S., Bunge, J., Leslin, C., Jeon, S. & Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *Isme J* **3**, 1365-1373.

Huang, L. N., Zhu, S., Zhou, H. & Qu, L. H. (2005). Molecular phylogenetic diversity of bacteria associated with the leachate of a closed municipal solid waste landfill. *Fems Microbiology Letters* **242**, 297-303.

Huson, D. H. & Mitra, S. (2012). Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN. *Evolutionary Genomics: Statistical and Computational Methods, Vol 2* **856**, 415-429.

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21**, 1552-1560.

Kakirde, K. S., Parsley, L. C. & Liles, M. R. (2010). Size does matter: Application-driven approaches for soil metagenomics. *Soil Biology & Biochemistry* **42**, 1911-1923.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C. & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16213-16216.

Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research* **12**, 656-664.

Krause, L., Diaz, N. N., Edwards, R. A., Gartemann, K. H., Kromeke, H., Neuweger, H., Puhler, A., Runte, K. J., Schluter, A., Stoye, J., Szczepanowski, R., Tauch, A. & Goesmann, A. (2008). Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. *Journal of Biotechnology* **136**, 91-101.

Krober, M., Bekel, T., Diaz, N. N., Goesmann, A., Jaenicke, S., Krause, L., Miller, D., Runte, K. J., Viehover, P., Puhler, A. & Schluter, A. (2009). Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. *Journal of Biotechnology* **142**, 38-49.

Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**, 118-123.

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**,

Lauro, F. M., DeMaere, M. Z., Yau, S., Brown, M. V., Ng, C., Wilkins, D., Raftery, M. J., Gibson, J. A. E., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J. M., Thomas, T. & Cavicchioli, R. (2011). An integrative study of a meromictic lake ecosystem in Antarctica. *Isme Journal* **5**, 879-895.

Lehembre, F., Doillon, D., David, E., Perrotto, S., Baude, J., Foulon, J., Harfouche, L., Vallon, L., Poulain, J., Da Silva, C., Wincker, P., Oger-Desfeux, C., Richaud, P., Colpaert, J. V., Chalot, M., Fraissinet-Tachet, L., Blaudez, D. and Marmeisse, R. (2013). Soil metatranscriptomics for mining eukaryotic heavy metal resistance genes. *Environmental Microbiology* doi: 10.1111/1462-2920.12143.

Leschine, S. B. (1995). Cellulose Degradation in Anaerobic Environments. *Annual Review of Microbiology* **49**, 399-426.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li, L. L., McCorkle, S. R., Monchy, S., Taghavi, S. & van der Lelie, D. (2009). Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnology for Biofuels* **2**,

Liu, J. A., Liu, W. D., Zhao, X. L., Shen, W. J., Cao, H. & Cui, Z. L. (2011). Cloning and functional characterization of a novel endo-beta-1,4-glucanase gene from a soil-derived metagenomic library. *Applied Microbiology and Biotechnology* **89**, 1083-1092.

Lockhart, R. J., Van Dyke, M. I., Beadle, I. R., Humphreys, P. & McCarthy, A. J. (2006). Molecular biological detection of anaerobic gut fungi (Neocallimastigales) from landfill sites. *Applied and Environmental Microbiology* **72**, 5659-5661.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* **30**, 434-+.

Loy, A., Arnold, R., Tischler, P., Rattei, T., Wagner, M. & Horn, M. (2008). probeCheck - a central resource for evaluating oligonucleotide probe coverage and specificity. *Environmental Microbiology* **10**, 2894-2898.

Luo, C. W., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. (2012). Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *Plos One* **7**,

Luo, Q. W., Krumholz, L. R., Najar, F. Z., Peacock, A. D., Roe, B. A., White, D. C. & Elshahed, M. S. (2005). Diversity of the microeukaryotic community in sulfide-rich zodletone spring (Oklahoma). *Applied and Environmental Microbiology* **71**, 6175-6184.

Lynd, L. R., Van Zyl, W. H., McBride, J. E. & Laser, M. (2005). Consolidated bioprocessing of cellulosic biomass: An update. *Current Opinion in Biotechnology* **16**, 577-583.

Lynd, L. R., Weimer, P. J., van Zyl, W. H. & Pretorius, I. S. (2002). Microbial cellulose utilization: Fundamentals and biotechnology. *Microbiology and Molecular Biology Reviews* **66**, 506-+.

Mader, U., Nicolas, P., Richard, H., Bessieres, P. & Aymerich, S. (2011). Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. *Current Opinion in Biotechnology* **22**, 32-41.

Magoc, T. & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. (2012). PANDAseq: PAired-eND Assembler for Illumina sequences. *Bmc Bioinformatics* **13**,

Martinez, D., Berka, R. M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S. E., Chapman, J., Chertkov, O., Coutinho, P. M., Cullen, D., Danchin, E. G. J., Grigoriev, I. V., Harris, P., Jackson, M., Kubicek, C. P., Han, C. S., Ho, I., Larrondo, L. F., de Leon, A. L., Magnuson, J. K., Merino, S., Misra, M., Nelson, B., Putnam, N., Robbertse, B., Salamov, A. A., Schmoll, M., Terry, A., Thayer, N., Westerholm-

Parvinen, A., Schoch, C. L., Yao, J., Barbote, R., Nelson, M. A., Detter, C., Bruce, D., Kuske, C. R., Xie, G., Richardson, P., Rokhsar, D. S., Lucas, S. M., Rubin, E. M., Dunn-Coleman, N., Ward, M. & Brettin, T. S. (2008). Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). *Nature Biotechnology* **26**, 553-560.

McCarren, J., Becker, J. W., Repeta, D. J., Shi, Y. M., Young, C. R., Malmstrom, R. R., Chisholm, S. W. & DeLong, E. F. (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16420-16427.

McDonald, J. E., Houghton, J. N. I., Rooks, D. J., Allison, H. E. & McCarthy, A. J. (2012). The microbial ecology of anaerobic cellulose degradation in municipal waste landfill sites: evidence of a role for fibrobacters. *Environmental Microbiology* **14**, 1077-1087.

McDonald, J. E., Allison, H. E. & McCarthy, A. J. (2010). Composition of the Landfill Microbial Community as Determined by Application of Domain- and Group-Specific 16S and 18S rRNA-Targeted Oligonucleotide Probes. *Applied and Environmental Microbiology* **76**, 1301-1306.

McDonald, J. E., de Menezes, A. B., Allison, H. E. & McCarthy, A. J. (2009). Molecular Biological Detection and Quantification of Novel Fibrobacter Populations in Freshwater Lakes. *Applied and Environmental Microbiology* **75**, 5148-5152.

McDonald, J. E., Lockhart, R. J., Cox, M. J., Allison, H. E. & McCarthy, A. J. (2008). Detection of novel Fibrobacter populations in landfill sites and determination of their relative abundance via quantitative PCR. *Environmental Microbiology* **10**, 1310-1319.

Mello, L. V., Chen, X. & Rigden, D. J. (2010). Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *Febs Letters* **584**, 2421-2426.

Mettel, C., Kim, Y., Shrestha, P. M. & Liesack, W. (2010). Extraction of mRNA from Soil. *Applied and Environmental Microbiology* **76**, 5995-6000.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. & Edwards, R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *Bmc Bioinformatics* **9**,

Miller, J. R., Koren, S. & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327.

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* **40**,

Neufeld, J. D., Schafer, H., Cox, M. J., Boden, R., McDonald, I. R. & Murrell, J. C. (2007). Stable-isotope probing implicates Methylophaga spp and novel Gammaproteobacteria in marine methanol and methylamine metabolism. *Isme Journal* **1**, 480-491.

Nicol, G. W., Campbell, C. D., Chapman, S. J. & Prosser, J. I. (2007). Afforestation of moorland leads to changes in crenarchaeal community structure. *Fems Microbiology Ecology* **60**, 51-59.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q. D., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. A., Ayvaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, V., Brooks, J. P., Buck, G. A., Buhay, C. J., Busam, D. A., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S. G., Chen, I. M. A., Chen, L., Chhibba, S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. A., Davidovics, N. J., Davis, C. C., DeSantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Dunne, W. M., Durkin, A. S., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor, A. A., Forney, L. J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H. Y., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Haake, S. K., Haas, B. J., Hamilton, H. A., Harris, E. L., Hepburn, T. A., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang, K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H. Y., Jordan, C., Joshi, V., Katancik, J. A., Keitel, W. A., Kelley, S. T., Kells, C., King, N. B., Knights, D., Kong, H. D. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., La Rosa, P. S., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C. C., Lozupone, C. A., Lunsford, R. D., Madden, T., Mahurkar, A. A., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavromatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. A., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Pop, M., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera, M. C., Rodriguez-Mueller, B., Rogers, Y. H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, P., Sathirapongsasuti, J. F., Schloss, J. A., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata, N., Segre, J. A., Shannon, W. D., Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. A., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. A., Walker, J., Wang, L., Wang, Z. Y., Ward, D. V., Warren, W., Watson, M. A., Wellington, C., Wetterstrand, K. A., White, J. R., Wilczek-Boney, K., Wu, Y. Q., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y. Z., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y. J., Zhu, Y. M., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. A., Highlander, S. K., Methe, B. A., Nelson, K. E., Petrosino, J. F., Weinstock, G. M., Wilson, R. K., White, O. & Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214.

Ogura, A., Lin, M., Shigenobu, Y., Fujiwara, A., Ikeo, K. & Nagai, S. (2011). Effective gene collection from the metatranscriptome of marine microorganisms. *Bmc Genomics* **12**,

Orsi, W., Biddle, J. F. & Edgcomb, V. (2013). Deep Sequencing of Subseafloor Eukaryotic rRNA Reveals Active Fungi across Marine Subsurface Provinces. *Plos One* **8**,

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. & Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* **33**, 5691-5702.

Palackal, N., Lyon, C. S., Zaidi, S., Luginbuhl, P., Dupree, P., Goubet, F., Macomber, J. L., Short, J. M., Hazlewood, G. P., Robertson, D. E. & Steer, B. A. (2007). A multifunctional hybrid glycosyl hydrolase discovered in an uncultured microbial consortium from ruminant gut. *Applied Microbiology and Biotechnology* **74**, 113-124.

Prakash, T. & Taylor, T. D. (2012). Functional assignment of metagenomic data: challenges and applications. *Briefings in Bioinformatics* **13**, 711-727.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J. & Gloeckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188-7196.

Porteous, L. A., Seidler, R. J. & Watrud, L. S. (1997). An improved method for purifying DNA from soil for polymerase chain reaction amplification and molecular ecology applications. *Molecular Ecology* **6**, 787-791.

Radax, R., Rattei, T., Lanzen, A., Bayer, C., Rapp, H. T., Urich, T. & Schleper, C. (2012). Metatranscriptomics of the marine sponge Geodia barretti: tackling phylogeny and function of its microbial community. *Environmental Microbiology* **14**, 1308-1324.

Rajendhran, J. & Gunasekaran, P. (2008). Strategies for accessing soil metagenome for desired applications. *Biotechnology Advances* **26**, 576-590.

Ransom-Jones, E., Jones, D. L., McCarthy, A. J. & McDonald, J. E. (2012). The Fibrobacteres: an Important Phylum of Cellulose-Degrading Bacteria. *Microbial Ecology* **63**, 267-281.

Rashamuse, K. J., Visser, D. F., Hennessy, F., Kemp, J., Roux-van der Merwe, M. P., Badenhorst, J., Ronneburg, T., Francis-Pope, R. & Brady, D. (2013). Characterisation of Two Bifunctional Cellulase-Xylanase Enzymes Isolated from a Bovine Rumen Metagenome Library. *Current Microbiology* **66**, 145-151.

Rieder, S. R. & Frey, B. (2013). Methyl-mercury affects microbial activity and biomass, bacterial community structure but rarely the fungal community structure. *Soil Biology & Biochemistry* **64**, 164-173.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P.,

Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P. & Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437.

Roesch, L. F., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K. M., Kent, A. D., Daroub, S. H., Camargo, F. A. O., Farmerie, W. G. & Triplett, E. W. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *Isme Journal* **1**, 283-290.

Rosewarne, C. P., Pope, P. B., Denman, S. E., McSweeney, C. S., O'Cuiv, P. & Morrison, M. (2011). High-Yield and Phylogenetically Robust Methods of DNA Recovery for Analysis of Microbial Biofilms Adherent to Plant Biomass in the Herbivore Gut. *Microbial Ecology* **61**, 448-454.

Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. (2012). Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *Plos One* **7**,

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA Sequencing With Chain-terminating Inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467.

Schluter, A., Bekel, T., Diaz, N. N., Dondrup, M., Eichenlaub, R., Gartemann, K. H., Krahn, I., Krause, L., Kromeke, H., Kruse, O., Mussgnug, J. H., Neuweger, H., Niehaus, K., Puhler, A., Runte, K. J., Szczepanowski, R., Tauch, A., Tilker, A., Viehover, P. & Goesmann, A. (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *Journal of Biotechnology* **136**, 77-90.

Schmieder, R., Lim, Y. W. & Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* **28**, 433-435.

Schmieder, R. & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864.

Schnoor, J. L. (2011). Cellulosic biofuels disappoint. *Environmental science & technology* **45**, 7099.

Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**,

Shi, Y., Tyson, G. W., Eppley, J. M. & DeLong, E. F. (2011). Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *Isme Journal* **5**, 999-1013.

Shin, S. G., Lee, C. S., Hwang, K., Ahn, J. H. & Hwang, S. (2008). Use of order-specific primers to investigate the methanogenic diversity in acetate enrichment system. *Journal of Industrial Microbiology & Biotechnology* **35**, 1345-1352.

Smith, C. J. & Osborn, A. M. (2009). Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *Fems Microbiology Ecology* **67**, 6-20.

Smith, C. J., Nedwell, D. B., Dong, L. F. & Osborn, A. M. (2006). Evaluation of quantitative polymerase chain reaction-based approaches for determining gene copy and gene transcript numbers in environmental samples. *Environmental Microbiology* **8**, 804-815.

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M. & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12115-12120.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. & Birney, E. (2002). The bioperl toolkit: Perl modules for the life sciences. *Genome Research* **12**, 1611-1618.

Stewart, F. J., Ulloa, O. & Delong, E. F. (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environmental microbiology* **14**, 23-40.

Stewart, F. J., Ottesen, E. A. & DeLong, E. F. (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *Isme Journal* **4**, 896-907.

Suzuki, M. T., Taylor, L. T. & DeLong, E. F. (2000). Quantitative Analysis of Small-Subunit rRNA Genes in Mixed Microbial Populations via 5'-Nuclease Assays. *Appl. Environ. Microbiol.* **66**, 4605-4614.

Tajima, K., Aminov, R. I., Nagamine, T., Matsui, H., Nakamura, M. & Benno, Y. (2001). Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR. *Applied and Environmental Microbiology* **67**, 2766-2774.

Takasaki, K., Miura, T., Kanno, M., Tamaki, H., Hanada, S., Kamagata, Y. & Kimura, N. (2013). Discovery of Glycoside Hydrolase Enzymes in an AvicelAdapted Forest Soil Fungal Community by a Metatranscriptomic Approach. *Plos One* **8**,

Teeling, H. & Gloeckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis-a bioinformatic perspective. *Briefings in Bioinformatics* **13**, 728-742.

Thomas, T., Gilbert, J. & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation* **2**,

Tripp, H. J., Hewson, I., Boyarsky, S., Stuart, J. M. & Zehr, J. P. (2011). Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Research* **39**,

Tveit, A., Schwacke, R., Svenning, M. M. & Urich, T. (2013). Organic carbon transformations in high-Arctic peat soils: key functions and microorganisms. *The ISME journal* **7**, 299-311.

Urich, T., Lanzen, A., Qi, J., Huson, D. H., Schleper, C. & Schuster, S. C. (2008). Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. *PLoS One* **3**,

Van Dyke, M. I. & McCarthy, A. J. (2002). Molecular biological detection and characterization of Clostridium populations in municipal landfill sites. *Applied and Environmental Microbiology* **68**, 2049-2053.

Wang, Y., Sheng, H.-F., He, Y., Wu, J.-Y., Jiang, Y.-X., Tam, N. F.-Y. & Zhou, H.-W. (2012). Comparison of the Levels of Bacterial Diversity in Freshwater, Intertidal Wetland, and Marine Sediments by Using Millions of Illumina Tags. *Applied and Environmental Microbiology* **78**, 8264-8271.

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., McHardy, A. C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S. G., Podar, M., Martin, H. G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov, A., Barry, K., Mikhailova, N., Kyrpides, N. C., Matson, E. G., Ottesen, E. A., Zhang, X. N., Hernandez, M., Murillo, C., Acosta, L. G., Rigoutsos, I., Tamayo, G., Green, B. D., Chang, C., Rubin, E. M., Mathur, E. J., Robertson, D. E., Hugenholtz, P. & Leadbetter, J. R. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-U517.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *Fems Microbiology Reviews* **28**, 127-181.
Westlake, K., Archer, D. B. & Boone, D. R. (1995). Diversity of Cellulolytic Bacteria in Landfill. *Journal of Applied Bacteriology* **79**, 73-78.

White, T. J., Bruns, T., Lee, S. & Taylor, J. (1990). Amplification and Direct Sequencing of Fungal Ribosomal RNA Genes for Phylogenetics. *Innis, M. a., Et Al. (Ed.). Pcr Protocols: a Guide to Methods and Applications. Xviii+482p. Academic Press, Inc.: San Diego, California, USA; London, England, Uk. Illus* 315-322.

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6578-6583. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6578-6583.

Wilson, D. B. (2009). Evidence for a novel mechanism of microbial cellulose degradation. *Cellulose* **16**, 723-727.

Wood, T. M. (1988). Preparation of Crystalline, Amorphous, and Dyed Cellulose Substrates. *Methods in Enzymology* **160**, 19-25.

Wooley, J. C., Godzik, A. & Friedberg, I. (2010). A Primer on Metagenomics. *Plos Computational Biology* **6**,

Xia, Y., Ju, F., Fang, H. H. P. & Zhang, T. (2013). Mining of Novel Thermo-Stable Cellulolytic Genes from a Thermophilic Cellulose-Degrading Consortium by Metagenomics. *Plos One* **8**,

Xiong, X., Frank, D. N., Robertson, C. E., Hung, S. S., Markle, J., Canty, A. J., McCoy, K. D., Macpherson, A. J., Poussier, P., Danska, J. S. & Parkinson, J. (2012). Generation and Analysis of a Mouse Intestinal Metatranscriptome through Illumina Based RNA-Sequencing. *Plos One* **7**,

Ying, J.-Y., Zhang, L.-M., Wei, W.-X. & He, J.-Z. (2013). Effects of land utilization patterns on soil microbial communities in an acid red soil based on DNA and PLFA analyses. *Journal of Soils and Sediments* **13**, 1223-1231.

Yu, K. & Zhang, T. (2012). Metagenomic and Metatranscriptomic Analysis of Microbial Community Structure and Gene Expression of Activated Sludge. *Plos One* **7**,

Yu, Z. T. & Morrison, M. (2004). Improved extraction of PCR-quality community DNA from digesta and fecal samples. *Biotechniques* **36**, 808-+.

Zamorano, M., Perez, J. I. P., Paves, I. A. & Ridao, A. R. (2007). Study of the energy potential of the biogas produced by an urban waste landfill in Southern Spain. *Renewable & Sustainable Energy Reviews* **11**, 909-922.

Zerbino, D. R. & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**, 821-829.

Zifcakova, L. & Baldrian, P. (2012). Fungal polysaccharide monooxygenases: new players in the decomposition of cellulose. *Fungal Ecology* **5**, 481-489.
Zhou, J. Z., Bruns, M. A. & Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62**, 316-322.

Zhu, W., Lomsadze, A. & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research* **38**,

Zumsteg, A., Luster, J., Goeransson, H., Smittenberg, R. H., Brunner, I., Bernasconi, S. M., Zeyer, J. & Frey, B. (2012). Bacterial, Archaeal and Fungal Succession in the Forefield of a Receding Glacier. *Microbial Ecology* **63**, 552-564.